Text-only Domain Adaptation for Low-Resource ASR Using Large Language Models

William Lamb, Dongge Han, Ondřej Klejch, Peter Bell
University of Edinburgh, Microsoft Corporation UK, University of Edinburgh, University of Edinburgh
{w.lamb, o.klejch, p.bell}@ed.ac.uk, dongge.han@microsoft.com

Abstract

Automatic Speech Recognition (ASR) increasingly mediates access to broadcast media, public discourse and cultural archives. For minoritised languages, however, the development of robust ASR systems is constrained by limited and domain-restricted text data. This paper investigates cross-lingual text expansion (XLTE), a method that uses a Large Language Model (LLM) to generate in-domain text in a low-resource language from high-resource language summaries. We further examine whether supervised fine-tuning on a small set of human-authored texts enhances generation quality. Using Scottish Gaelic as a case study, we show that synthetic text generated via fine-tuned XLTE can be used to train an external language model that reduces Word Error Rate (WER) by 24.48% in a previously unseen broadcast domain. Our findings demonstrate that text-only domain adaptation through cross-lingual generation can strengthen speech technology in sparse data settings. Beyond engineering gains, the approach offers a scalable pathway for improving the digital representation, accessibility and sustainability of minoritised-language media and cultural heritage.

Keywords: speech recognition, domain adaptation, low-resource languages, Scottish Gaelic, large language models

1. Introduction

A central challenge in improving linguistic diversity in modern automatic speech recognition (ASR) systems is data sparsity. Although contemporary end-to-end ASR architectures jointly model acoustic and linguistic information, a strong external language model (LM) remains important in low-resource settings, where it can partially compensate for weaker acoustic representations (Wallington et al., 2021). For many minoritised languages, however, the difficulty lies not only in the limited quantity of available text but also in its narrow domain coverage.

Scottish Gaelic illustrates this problem clearly. The language is spoken by $69{,}700$ individuals in Scotland (1.3% of the population) (National Records of Scotland, 2022). Publicly available Gaelic text is dominated by BBC regional radio news scripts and is heavily skewed towards local reporting. When such data are used to train an LM for related but distinct domains, for example televised national or international news, performance deteriorates because of domain mismatch. This is not merely a technical inconvenience; it constrains the development of speech technologies that enable access to contemporary media, public discourse and cultural archives in minoritised languages.

Addressing domain mismatch is therefore central to broader efforts within the Social Sciences and Humanities (SSH) to build sustainable and equitable language technology infrastructures. ASR systems increasingly mediate access to minority-language media and oral heritage, yet creating sufficiently broad and representative training corpora remains prohibitively expensive for most low-resource languages.[1]

Ideally, an external LM for ASR would be trained on large volumes of in-domain transcribed text. In practice, such data rarely exist and using multilingual large language models (LLMs) presents one alternative to finding or creating genuine supervised speech data. If a target language appears in an LLM's pre-training data, it may be prompted to synthesise new text in that language using zero-shot or few-shot methods. Previous work has shown that fine-tuning can further improve generation quality when in-domain examples are available (Bell et al., 2021). However, in genuinely low-resource settings, even small quantities of domain-specific data may be unavailable.

Recent studies have explored synthetic data augmentation for low-resource languages (Lucas et al., 2024; Samuel et al., 2024; Alcoba Inciarte et al., 2024). Yet LLM outputs for such languages often exhibit structural interference from dominant high-resource languages, typically English (Robinson et al., 2023; Lai et al., 2023). This may reflect English-centric representational biases in multilingual models (Papadimitriou et al., 2023; Wendler et al., 2024). For ASR domain adaptation, however, perfectly fluent output is not required. Theoretically, improvements could come from synthetic text simply increasing coverage of n-grams present in evaluation data.

[1] Crowdsourcing is one option. See 'Opening the Well', a community-driven and ASR-assisted Gaelic folklore transcription initiative: https://fosgladh.tobarandualchais.co.uk/en.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 95–102
11 May 2026.

In this study we simulate a scenario where limited LM training data exist for a low-resource language and propose a solution based on **cross-lingual text expansion (XLTE)**. XLTE generates in-domain text in a low-resource target language by prompting an LLM with summaries in a high-resource source language. The method combines cross-lingual transfer with controlled expansion: a short source-language summary is transformed into a longer target-language text aligned with the desired domain.

Beyond zero-shot XLTE, we introduce a supervised fine-tuning approach in which the LLM is trained to reconstruct original low-resource texts from summaries written in a high-resource language. Unlike prior work that improves n-gram LMs through, for example, transformer-based rescoring (Wang et al., 2019), our approach synthesises entirely new in-domain text in the target low-resource language.

For practical reasons, we fine-tune OpenAI's GPT-4o model,[2] adapting it to map English summaries of Gaelic regional news texts to their original Gaelic counterparts. We then prompt the fine-tuned model with wide-domain English summaries from the CNN news dataset (Hermann et al., 2015) to generate international news texts in Gaelic. These synthetic texts are then used to train an n-gram-based external LM, which is evaluated in a downstream ASR task involving the Gaelic television news programme *An Là* ('The Day').

Our central objective is to inject world knowledge encoded in a high-resource language, English, into a low-resource Gaelic training corpus through controlled cross-lingual generation. We evaluate zero-shot XLTE versus supervised fine-tuning, XLTE versus machine translation – as a data augmentation strategy – and the impact of synthetic data on downstream ASR performance. Building on our earlier introduction of XLTE (Lamb et al., 2025), which focused on intrinsic evaluation of synthetic narrative data, this paper shifts the emphasis to domain adaptation and extrinsic validation, demonstrating measurable reductions in WER in a low-resource ASR task.

Summary of contributions:

1. We evaluate an LLM-based cross-lingual data augmentation strategy for domain adaptation in low-resource ASR.
2. We demonstrate that supervised fine-tuning improves synthetic text quality over a baseline LLM across intrinsic evaluation metrics.
3. We achieve substantial reductions in Word Error Rate (WER) for a previously unseen domain in a real-world Gaelic ASR task.

[2]Model version: gpt-4o-2024-05-13.

2. Methodology

2.1. Text generation

XLTE requires aligned pairs of English summaries and Gaelic texts. To construct these, we prompted the baseline GPT-4o model to produce English summaries of the original Gaelic news articles (see Table 3) using the instruction: *Your role is to summarise the given news story in 3 to 4 sentences in English.* These summaries serve as the source-language stimuli for subsequent expansion. The remaining stages of the pipeline are illustrated in Figure 1 and detailed below.

We then tested whether supervised fine-tuning improves performance relative to the baseline GPT-4o model. For fine-tuning, the model was trained with the instruction: *You will receive a summary of a news story in English. Expand the summary into a much longer news story in Scottish Gaelic.* Prompting in Gaelic led to reduced performance and, therefore, was not pursued further.

At generation time, we expanded English in-domain summaries into corresponding Gaelic texts. Depending on the experimental condition, we generated between 100 and 2,701 synthetic news texts.

2.2. Evaluation metrics

We employ intrinsic metrics as proxies for downstream ASR performance: Mean Word Count (MWC), English-to-Gaelic ratio (en:gd), Neologism Ratio (Neo), Perplexity (PPL) and Self-BLEU (SB). Lower values are preferred for all metrics except mean word count. Each metric captures a different property of the generated text.

Mean Word Count (MWC) measures average volubility per generated sample. Higher MWC implies that fewer API calls are required to reach a target corpus size, thereby reducing generation cost.

The English-to-Gaelic ratio (en:gd) estimates language uniformity. It is computed by dividing the number of tokens matching a large English dictionary by those matching a large Gaelic dictionary. Tokens absent from both dictionaries are classified as neologisms, which include hallucinated forms and other out-of-vocabulary items. The Neologism Ratio (Neo) reports the proportion of such tokens.

Perplexity (PPL) measures the predictive fit between a language model and a text sample. Although widely used as a proxy for language model quality, it correlates imperfectly with human judgements (Stureborg et al., 2024) and is sensitive to punctuation (Wang et al., 2023) and length effects (Meister and Cotterell, 2021). For consistency, we lowercase and normalise all texts, to-

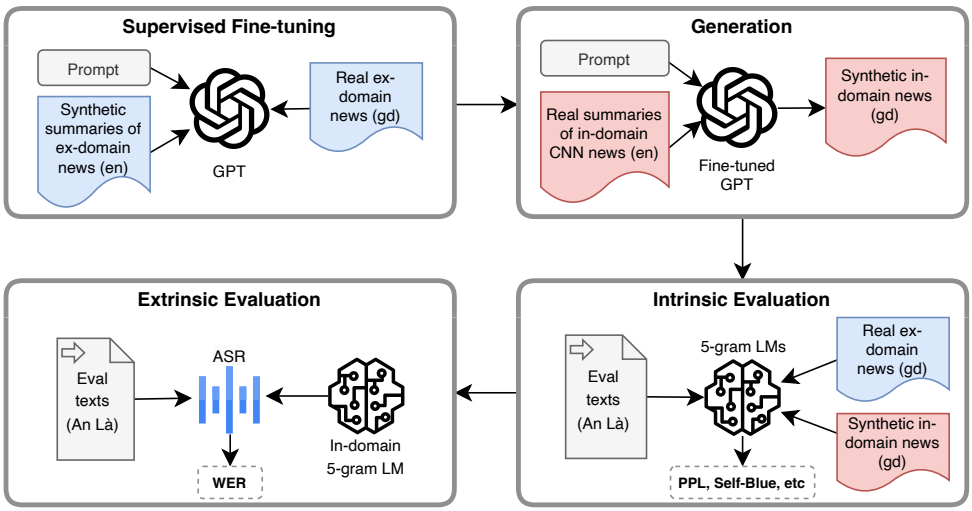

Figure 1: Training and evaluation pipeline (en 'English'; gd 'Gaelic'; PPL 'perplexity')

kenise using a byte-pair encoding model (Sennrich et al., 2016) trained on a dataset of Gaelic news scripts, and compute perplexity with 5-gram language models employing modified Kneser-Ney smoothing (KenLM (Heafield et al., 2013)).

Human-authored texts typically show greater lexical diversity than synthetic texts (Yu et al., 2024). To assess our texts' lexical diversity, we compute Self-BLEU (SB), which measures the average n-gram overlap between each sentence in a text and all others (Zhu et al., 2018). Lower SB indicates greater diversity. For a set of sentences $\{s_1, s_2, \ldots, s_m\}$, Self-BLEU is defined as:

$$\mathrm{SB} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{BLEU}\big(s_i, \{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_m\}\big) \quad (1)$$

For extrinsic evaluation, we use the best-performing fine-tuned GPT-4o model to generate a larger synthetic corpus. An n-gram LM is trained on this material and integrated into a low-resource ASR system. We then compare Word Error Rate (WER) against a baseline LM trained solely on the available training data (Tr). The key question is whether incorporating synthetic text yields measurable WER reduction.

3. Datasets

With the permission of BBC ALBA, we scraped our out-of-domain data from texts posted on the BBC's *Naidheachdan* 'News' website.[3] These originated as scripts for news programmes on the *Radio nan Gàidheal* radio station. Although the original radio programmes cover regional, national and international stories, the *Naidheachdan* web texts are heavily skewed towards regional reporting. We scraped approximately 37,000 bulletins (c. 6.5M words) spanning January 2012 to October 2023. After de-duplication, we split the corpus into three disjoint RnG subsets: a training set (Tr) and validation set (Val) for fine-tuning, and a generation set (Gen) used to create English summary stimuli using GPT-4o (see Section 2). We treat the BBC RnG scripts as the source domain and BBC ALBA's *An Là* as the target domain.

Our evaluation set (Eval) was drawn from BBC ALBA's *An Là* ('The Day') news programme[4] and comprises verbatim transcripts of six 30-minute episodes with aligned audio. We use Eval both for intrinsic evaluation (perplexity) and for extrinsic evaluation, where we report Word Error Rate (WER) on the corresponding audio.

The raw CNN data were from the training subset of the `cnn_dailymail` dataset (Hermann et al., 2015), available at `HuggingFace`.[5] The CNN news articles (all English-medium) originally appeared between April 2007 and April 2015. We deployed two separate generation subsets from these data: Gen1, comprising 1000 GPT-4o-generated summaries of CNN articles used during intrinsic evaluation; and Gen2, a larger, disjoint 2701 article dataset, which we used during extrinsic evaluation. Because Gen2 is used only as a stimulus source for generation, leveraging the dataset highlights rather than model-generated summaries does not affect the downstream comparison. Table 3 summarises these datasets.

[3] https://www.bbc.co.uk/naidheachdan

[4] https://www.bbc.co.uk/programmes/b00drynf

[5] https://huggingface.co/datasets/abisee/cnn_dailymail

Source	Lang	Subset	N	Words (gd)
RnG	gd	Tr	1000	240,866
RnG	gd	Val	200	40,729
RnG	gd	Gen	2000	444,953
An Là	gd	Eval	7	24,725
CNN	en	Gen1	1000	n/a
CNN	en	Gen2	2701	n/a

Table 1: Datasets (Key: RnG 'Radio nan Gàidheal'; Lang 'original language'; gd 'Gaelic'; en 'English'; Tr = fine-tuning training set; Val = validation set; Gen = summary stimuli; Eval = ASR evaluation set; N = number of documents.)

Input	PPL	SB	MWC	en:gd	Neo
Real (RnG)	**317.3**	0.41	226.1	0.02	0.03
Gen (RnG)	422.6	0.47	381.0	0.02	0.02
Gen (CNN)	494.5	**0.40**	377.7	0.03	0.04

Table 2: GPT-4o baseline: Intrinsic evaluation metrics for generated texts derived from RnG-Gen and CNN-Gen1 summaries, and ground-truth texts from the RnG-Gen set ($N = 400$). Key: PPL = perplexity; SB = Self-BLEU; MWC = mean word count; en:gd = English-to-Gaelic token ratio; Neo = proportion of out-of-dictionary tokens.

4. Experimental Results

4.1. Establishing Baseline Performance

We began by summarising the texts in English using the baseline GPT-4o model to produce the required summary-text pairs.[6] Then, we generated 400 texts using the RnG-Gen and CNN-Gen1 summaries and compared the synthesised texts to real ones from the RnG-Gen set. As seen in Table 2, when using the GPT-4o-base model, LMs built from the synthetic data obtained higher PPL values against the Eval set than an LM built from the real data. So, the real data produce an LM with a better fit to the target domain. Although past research has found real text to be more diverse than synthetic text (Yu et al., 2024), texts generated from CNN-Gen1 show a slightly lower Self-BLEU score than the real data and those generated from RnG-Gen. This textual diversity could be advantageous to the downstream ASR task. Yet, the English to Gaelic and Neologism ratios are higher than the real and generated RnG texts. The international focus of the CNN summaries may dispose the LLM to output more English and out-of-dictionary tokens.

[6]Summarisation hyperparameter settings: n=1, τ=1, top-p=0.85, frequency penalty=0, presence penalty=0.2 and max tokens=250. The maximum token count translates to roughly 3 to 4 sentences in English.

Model	PPL	SB	MWC	en:gd	Neo
FT100	**283.8**	**0.25**	457.5	0.05	0.09
FT200	313.4	0.25	454.9	0.06	0.09
GPT-4o	494.5	0.40	379.2	0.03	0.05

Table 3: Fine-tuning experiments: Intrinsic evaluation metrics for fine-tuned models (FT100, FT200) and the GPT-4o baseline.

4.2. Effect of Supervised Fine-tuning

To investigate whether supervised fine-tuning improves text synthesis quality, we fine-tuned GPT-4o on 100, 200 and 400 summary-text pairs from the RnG training set. We refer to these models as FT100, FT200 and FT400. After a short hyperparameter study, we settled on fine-tuning for 3 epochs with a batch size of 1 and a learning rate multiplier of 2. Loss trends indicated effective early-stage learning with no signs of over-fitting.

We generated 400 texts from the CNN-Gen1 set using the GPT-4o-base, FT100 and FT200 models.[7] Table 3 compares them using the intrinsic evaluation metrics. The LM associated with FT100 had the lowest PPL value against the Eval set. This suggests that supervised fine-tuning boosts performance for this task over generating from the baseline GPT model. The FT100-based LM is also a better fit to the Eval set and more lexically diverse than an LM built from out-of-domain real data (cf. 'Real: RnG' in Table 2). Notably, given that the FT100 model's LM produces a lower PPL value than the FT200 model's LM, the ideal number of fine-tuning examples appears to be below 200 for these data and this use-case.

In general, Table 3 shows that, compared to the baseline model, the fine-tuned models produce more diverse text, a higher MWC, a higher proportion of English text and a higher proportion of neologisms. While the greater diversity and higher MWC are likely favourable for our task, the inflated English and neologism could be detrimental. Impressionistically, the baseline model produces text that is more coherent, but also more generic. The FT models, in comparison, are stylistically closer to the real news-scripts.

4.3. Translation versus Generation

To consider whether machine translation (MT) yields superior results to GPT-4o-based genera-

[7]We began using a top-p of 0.85, but reduced it to 0.7 after further testing. The other generation hyperparameters were: n=1, τ=1, frequency penalty=0.2; presence penalty=0.5 and max tokens=1000. Experimentation with the baseline GPT-4o model suggested that these settings provided good diversity, less repetition and the longer outputs required for our task.

Source	PPL	SB	MWC	en:gd	Neo
Gen	**381.9**	**0.43**	419.2	0.06	0.06
MT	505.1	0.45	640.7	0.05	0.05

Table 4: Generated vs. machine-translated text: Intrinsic evaluation metrics.

tion, we expanded 800 summaries from the CNN-Gen1 set using our FT100 model. In tandem, we translated English texts from the same dataset to Gaelic using Google Translate's API. After building n-gram language models from the generated and machine-translated text, we calculated the intrinsic evaluation metrics. To ensure fair comparisons between the MT and Generated (Gen) datasets, we controlled for word count.

Table 4 shows that an LM built from FT100-generated text (Gen) achieves lower perplexity on the Eval set than one built from the MT text. The lower Self-BLEU score indicates that the Gen text is also slightly more diverse than the MT text, despite its lower MWC. The stronger supervision signal associated with MT may explain the marginally lower en:gd and neologism ratios for the MT text. In sum – for this task, these models and these languages – XLTE produces higher quality and more diverse text than a state-of-the-art MT system, but shows marginally inflated English and neologism ratios.

Ultimately, improvements in intrinsic metrics are only meaningful if they translate into downstream ASR gains. We therefore evaluate whether synthetic data reduce WER under realistic decoding conditions.

4.4. Extrinsic Evaluation

We evaluated our approach in a real-world setting by generating a large volume of text with FT100, incorporating it into an external LM and using it for an ASR task within the target domain. The goal was to determine whether XLTE-derived text enhances WER, which would suggest effective domain adaptation.

The ASR system's acoustic model was our top-line Gaelic model at the time the research was carried out (details of architecture in Klejch et al., 2025[8]). Acoustic features were extracted from the 18th layer of an XLS-R 300M model (Babu et al., 2022), which underwent continual pre-training on Gaelic and English. The English text came from the MGB-1 corpus (Bell et al., 2015). The acoustic model employed a TDNN-F architecture with 1,000 BPE units and shared a tokeniser with the language model. Evaluation was conducted on the Eval set (175 mins of aligned acoustic and textual data) described in §3.

[8]Between the time that the research was conducted and the publication of Klejch et al., 2025, the WER on the *An Là* testset dropped from 14.81% to 10.4%.

Dataset	Train	Gen	Train+Gen	Top-line
gd	29.37	22.86	22.18	–
gd+en	21.01	19.30	**18.86**	14.81

Table 5: WER (%) for the ASR task using LMs built from real data (Train), synthetic data (Gen), and their combination (Train+Gen). Interpolation with English data is shown in the gd+en row.

Table 5 reports WER results for external LMs trained on three datasets: real data from Tr ('Train': 60,994 words), synthetic data generated from 2,701 Gen2 summaries using FT100 ('Gen': 1,038,793 words) and the concatenation of these two datasets ('Train+Gen': 1,099,787 words).

To simulate a low-resource setting, the real data were restricted to the 100 texts used to fine-tune FT100. To better model code-switching, which is common in spoken Gaelic (Smith-Christmas, 2012), we also trained variants that incorporated English data from the MGB-1 corpus (Bell et al., 2015).

For reference, Table 5 includes the WER of our current production-grade external LM ('Top-line') evaluated on the same test set. This model is trained on the full available Gaelic corpus and therefore represents an upper-bound benchmark rather than a system operating under the same low-resource constraints.

Using an LM built from 100 real Gaelic texts yields a WER of 29.37%, which drops to 21.01% when interpolating with English data. Using LMs built from the synthetic data further lowers the WER (22.86% for Gaelic only, 19.3% for Gaelic+English), validating our approach. Fine-tuning GPT-4o on just 100 summary-text pairs synthesises an effective, in-domain corpus over an order of magnitude larger than the original, achieving a 22.17% relative WER reduction. Combining the real and generated data brings the reduction to 24.48% ($29.37\%\rightarrow22.18\%$) and interpolating the English data increases the relative WER reduction to 35.78% ($29.37\%\rightarrow18.86\%$). The WER of 18.86% is only 4 percentage points away from the top-line on this task.

Typical of modern Gaelic speech, our Eval set contains many English words: roughly 17% of the total. Notably, our synthetic data have a higher en:gd ratio than the training data (0.8 vs 0.34). This, along with the fact that deletions account for the biggest reduction in WER between the training and generated data, led us to investigate whether our approach mainly enhances English speech recognition.

Table 6 shows that 66% of the total WER reduction (514 of 784 errors) can be attributed to improve-

Language	Subs	Ins	Dels	Sum	% Total
English	-282	49	-281	-514	66%
Gaelic	-214	-11	-45	-270	34%

Table 6: Change in error counts after replacing the external LM trained on real data with one incorporating synthetic data. Negative values indicate reductions in errors. The final column shows each language's proportion of the total reduction in errors ($n = 784$).

ments with English tokens, with 55% of these gains arising from fewer deletions. Although English insertions increase slightly ($n = 49$), the net error reduction remains strongly positive. Indeed, 34% of the total error reduction (270 errors) relates to Gaelic tokens, confirming that the synthetic data improve recognition of both languages. While gains for English are more pronounced, the reduction in Gaelic errors provides evidence that XLTE enhances in-domain language modelling; it does not merely boost English coverage. This is encouraging for future efforts applying XLTE to low-resource ASR settings (cf. Joshi and Singh, 2022).

5. Conclusions

This study demonstrates that fine-tuning an LLM on a modest set of human-authored texts can generate a substantially larger in-domain corpus for a low-resource language. Intrinsic evaluation shows that cross-lingual text expansion (XLTE) produces material better aligned with the target domain than machine translation, yielding lower perplexity and greater internal diversity. When used to train an external language model, the synthetic corpus delivers substantial downstream gains, reducing WER by up to 24.48% under low-resource conditions and 35.78% when interpolated with English data.

These results establish synthetic text generation as an effective mechanism for domain adaptation in low-resource ASR. Improvements are observed for both Gaelic and English tokens in a bilingual setting, indicating enhanced in-domain coverage rather than inflated majority-language recognition. The method is computationally lightweight, scalable and does not require additional parallel data, making it readily transferable to other minoritised languages and under-represented domains.

The study also highlights the continuing value of human-generated texts (Bird and Yibarbuk, 2024). Even relatively small, carefully curated human-produced datasets provide the structural and stylistic signal necessary for effective fine-tuning. In this sense, community-authored material functions as critical digital infrastructure for low-resource language technology.

Several limitations remain. Fine-tuning was conducted via a proprietary API, which constrains transparency. The growing availability of open-weight models makes replication with them feasible, although recent work suggests that open-weight models are generally less performative for Gaelic than leading proprietary ones (Devine et al., 2026). The extrinsic evaluation assumes a strong acoustic model; future work should examine stricter low-resource acoustic conditions. Finally, validation was limited to a single broadcast domain, and broader genre coverage will be required to assess generalisability.

Beyond engineering gains, the findings carry implications for the Social Sciences and Humanities. Speech recognition increasingly provides a conduit for searching and accessing broadcast media, folklore and oral history archives. For minoritised languages, domain mismatch in language modelling affects not only accuracy but cultural visibility. Strengthening domain-specific language models is therefore a prerequisite for equitable digital representation and sustainable access to linguistic heritage. In the Gaelic context, initiatives such as *Opening the Well* illustrate how advances in language technology can enhance the digital presence of cultural materials, while community engagement in turn strengthens the technological ecosystem, creating a mutually reinforcing cycle.

Statement on Ethics

Institutional ethical review was initiated on 2 March 2023 and approved on 13 March 2023 by the Ethics Officer of the host institution. The study involved no human participants and was assessed as presenting minimal risk. The authors nonetheless recognise the environmental implications associated with training and deploying large language models. While the primary computational costs arise during large-scale pre-training, fine-tuning and text generation also entail energy consumption. Because this study relied on a proprietary model accessed via API, precise estimates of compute usage and associated carbon emissions are not publicly available.

Code and Data Availability

All code and synthetic data generated by the fine-tuned GPT-4o model is available at `https://github.com/razorfish17/XLTE`. The BBC *Naidheachdan* corpus cannot be redistributed due to third-party licensing restrictions, but the scraping methodology, including the date range, URL structure and data selection criteria, is described in §3 to facilitate replication.

Acknowledgements

This work has benefited from the support of the Scottish Government (Grant name: 'Ecosystem for Interactive Speech Technologies'). Thanks to the anonymous peer reviewer for their helpful comments. Thanks also to Mr Reamonn Lenkas at BBC ALBA for his assistance gathering the *naidheachdan* data and Mr Cailean Gordon for transcribing the *An Là* episodes.

References

Alcides Alcoba Inciarte, Sang Yun Kwon, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2024. On the utility of pretraining language models on synthetic data. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 265–282, Bangkok, Thailand. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech*.

Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. 2021. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66.

Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, et al. 2015. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693. IEEE.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Peter Devine, William Lamb, Beatrice Alex, Ignatius Ezeani, Dawn Knight, Mícheál J. Ó Meachair, Paul Rayson, and Martin Wynne. 2026. GaelEval: Benchmarking LLM performance for Scottish Gaelic. In *Proceedings of LLMs4SSH*, Palma de Mallorca.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Raviraj Joshi and Anupam Singh. 2022. A simple baseline for domain adaptation in end to end ASR systems using synthetic data. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 244–249.

Ondřej Klejch, William Lamb, and Peter Bell. 2025. A Practitioner's Guide to Building ASR Models for Low-Resource Languages: A Case Study on Scottish Gaelic. In *Interspeech 2025*, pages 728–732.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189.

William Lamb, Dongge Han, Ondrej Klejch, Beatrice Alex, and Peter Bell. 2025. Synthesising a corpus of Gaelic traditional narrative with cross-lingual text expansion. In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 12–26, Abu Dhabi [Virtual Workshop]. International Committee on Computational Linguistics.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of Guarani-Spanish neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6385–6397.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. 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 5328–5339, Online. Association for Computational Linguistics.

National Records of Scotland. 2022. Scotland's census 2022. https://www.

scotlandscensus.gov.uk. Accessed: 2024-08-14.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can LLMs Augment Low-Resource Reading Comprehension Datasets? Opportunities and Challenges.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Cassie Smith-Christmas. 2012. *I've lost it here dè a bh'agam: Language shift, maintenance, and code-switching in a bilingual family*. Ph.D. thesis, University of Glasgow.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Electra Wallington, Benji Kershenbaum, Peter Bell, and Ondřej Klejch. 2021. On the learning dynamics of semi-supervised training for ASR. In *Interspeech 2021: The 22nd Annual Conference of the International Speech Communication Association*, pages 716–720. International Speech Communication Association.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2023. Perplexity from PLM is unreliable for evaluating text quality.

Yiren Wang, Hongzhao Huang, Zhe Liu, Yutong Pang, Yongqiang Wang, ChengXiang Zhai, and Fuchun Peng. 2019. Improving n-gram language models with pre-trained deep transformer.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.