Synthesising a Corpus of Gaelic Traditional Narrative with Cross-Lingual Text Expansion

William Lamb[1], Dongge Han[1,2], Ondřej Klejch[1], Beatrice Alex[1], Peter Bell[1]

[1]University of Edinburgh, [2]Microsoft Corporation
Correspondence: w.lamb@ed.ac.uk

Abstract

Advances in large language modelling have disproportionately benefited high-resource languages due to their vastly greater training data reserves. This paper proposes a novel cross-lingual text expansion (XLTE) technique using multilingual large language models (MLLMs) to mitigate data sparsity in low-resource languages. We apply XLTE to the domain of traditional Scottish Gaelic storytelling to generate a training corpus suitable for language modelling, for example as part of an automatic speech recognition system. The effectiveness of this technique is demonstrated using OpenAI's GPT-4o, with supervised fine-tuning (SFT) providing decreased neologism rates and a 57.2% reduction in perplexity over the baseline model. Despite these promising results, qualitative analyses reveal important stylistic divergences between synthesised and genuine data. Nevertheless, XLTE offers a promising, scalable method for synthesising training sets in other languages and domains, opening avenues for further improvements in low-resource language modelling.

1 Introduction

Recent breakthroughs in natural language processing (NLP), particularly the development of large language models (LLMs), have principally benefited well-resourced languages like English and Spanish. Most of the world's languages remain marginalised, however, due to a lack of suitable training data (Magueresse et al., 2020; Joshi et al., 2020). Training even a GPT-2-scale LLM, for example, requires roughly 10B tokens of text.[1] This far exceeds the extant corpus of most low-resource languages (LRLs). While many high-resource languages (HRLs) can exploit web-scale datasets, languages like Gaelic – spoken by 69,701 individuals in Scotland (National Records of Scotland, 2022) – can offer only a minute fraction of such data.

A promising solution to the sparsity problem facing LRLs is coupling synthetic text generation with cross-lingual transfer (Chen et al., 2019). Multilingual large language models (MLLMs) like BLOOM (Scao et al., 2022), GPT (Radford et al., 2018) and LLaMA (Dubey et al., 2024) are trained on manifold languages, enabling them to transfer knowledge from high-resource languages to tasks involving LRLs. What if we could harness these cross-lingual capacities to produce useful training data for LRLs, for instance towards language modelling? In other words, could we prompt a MLLM to generate a training corpus in a LRL?

The digital text available for Gaelic is approximately 150M tokens but the language is better-resourced for audio data. This is due, in part, to the thousands of hours of ethnographic recordings made of Gaelic speakers in the mid-20th century. If we could reliably transcribe these audio data, we could substantially augment the language's textual resources. A key objective of ongoing work is to automatically transcribe recordings of traditional narrative, many of which are hosted on the online portal Tobar an Dualchais / Kist o Riches.[2] About 1M words of high-quality narrative text exist from earlier digitisation and recognition efforts (Sinclair et al., 2022; Meaney et al., 2024), but a much larger corpus is needed to improve automatic speech recognition (ASR) for this domain (Evans et al., 2022).

We hypothesise that we can increase our narrative training data by deploying a novel *cross-lingual text expansion* (XLTE) method. Text expansion is converting a short text into a longer one (Dong et al., 2022). XLTE couples expansion with translation: it involves prompting a MLLM with a summary in one language to generate an extended

---

[1] Radford et al., 2018 used 40GB of text to train GPT-2, which amounts to about 10B tokens.

[2] `https://www.tobarandualchais.co.uk`

Proceedings of the 5th Celtic Language Technology Workshop (CLTW 5), pages 12–26
Abu Dhabi, UAE (Online), January 20, 2025

text in another language. For our use-case, we fine-tune a MLLM using transcriptions of oral Gaelic narratives paired with their English summaries. We then generate a synthetic corpus of narrative using a held out set of English summaries.[3] We hypothesise that fine-tuning a MLLM for this task will improve results over generating using a baseline model. We use OpenAI's GPT-4o model (OpenAI, 2024; Islam and Moushi, 2024)[4] but expect that XLTE could be extended to other MLLMs with API access and fine-tuning capabilities, such as Claude (Anthropic, 2024) or LLaMA, provided they offer similar coverage for a target LRL.

To explore our hypotheses, we adopt the following research questions:

1. What benefits, if any, does supervised fine-tuning (SFT) offer over the baseline GPT model for using XLTE to generate a synthetic corpus of Gaelic traditional narrative?

2. How do texts generated using XLTE compare with genuine ones across intrinsic evaluation metrics?

3. What stylistic differences, if any, can be detected between our synthetic and genuine narrative texts?

The organisation of the remaining paper is as follows: §2 provides background information and surveys relevant literature; §3 describes our datasets and methodology; §4 presents and discusses our results and, finally, §5 offers concluding remarks and future research possibilities.

2 Background and Related Work

2.1 Large Language Models (LLMs)

Progress in NLP has been catalysed by the emergence of large language models, especially the variety known as Generative Pre-trained Transformers (GPTs). During pre-training, these models use self-attention (Vaswani et al., 2017) within a next-token prediction task, inducing the relative importance of each token in an input stream to every other token (Raiaan et al., 2024). Through this process, they can compress a vast input corpus (e.g. all of the Internet's text) into a high-dimensional, context-cognisant representation of a language's vocabulary, linguistic features and associated 'world-knowledge' (Zhao et al., 2023).

For many applications, a pre-trained base-model can be improved through additional training known as *supervised fine-tuning* (SFT). SFT involves updating some of the model's parameters using labelled data, and biasing the model to produce more accurate classification or generation results given a particular prompt (e.g. 'Expand the given summary into a longer traditional narrative in Scottish Gaelic') and supervised dataset (Chen et al., 2024; Mosbach et al., 2023; Qin et al., 2022). SFT is quicker and requires fewer resources than training a model from scratch but delivers improved performance for many applications (Zhao et al., 2023). In the present study, we investigate whether SFT enhances the capabilities of GPT-4o to generate Gaelic traditional narrative texts over using the base model. In general, this study aligns with a body of NLP research that examines augmenting or creating domain-specific training data where little to none exist.

2.2 LLM-based Synthetic Text Generation

Using LLMs to synthesise training corpora and supervised data is a growing research area across multiple domains (see Ding et al., 2024, Guo and Chen, 2024 and Sufi, 2024 for recent reviews). Common use-cases include generating labelled medical data (Falis et al., 2024), plausible questionnaire responses (Hämäläinen et al., 2023), multi-turn dialogue data (Xu et al., 2023) and low-resource machine translation data (Lucas et al., 2024; Hong et al., 2024). Notably, one study on another LRL (Arabic) demonstrated that LLMs built on GPT-2-generated text performed comparably to ones built using the outputs of optical character recognition (OCR) and ASR for a range of natural language understanding tasks (Alcoba Inciarte et al., 2024). Evidence has emerged that training LLMs iteratively on synthetic text leads to diminished linguistic diversity and model collapse (Guo and Chen, 2024; Dohmatob et al., 2024; Shumailov et al., 2024). Nevertheless, this is not a concern for the present study given that we deploy first-generation synthesised data only; no iteration is involved.

Salient to our present aims, several recent papers have investigated how well LLMs produce long-form texts, such as fiction and storytelling (Yang et al., 2022; Xie and Riedl, 2024; Tian et al., 2024;

---

[3] Performance tends to be better when using English prompts versus ones in the target language (Bareiß et al., 2024)).

[4] OpenAI approached us in June 2023 as part of an initiative to collaborate with low-resource speech communities.

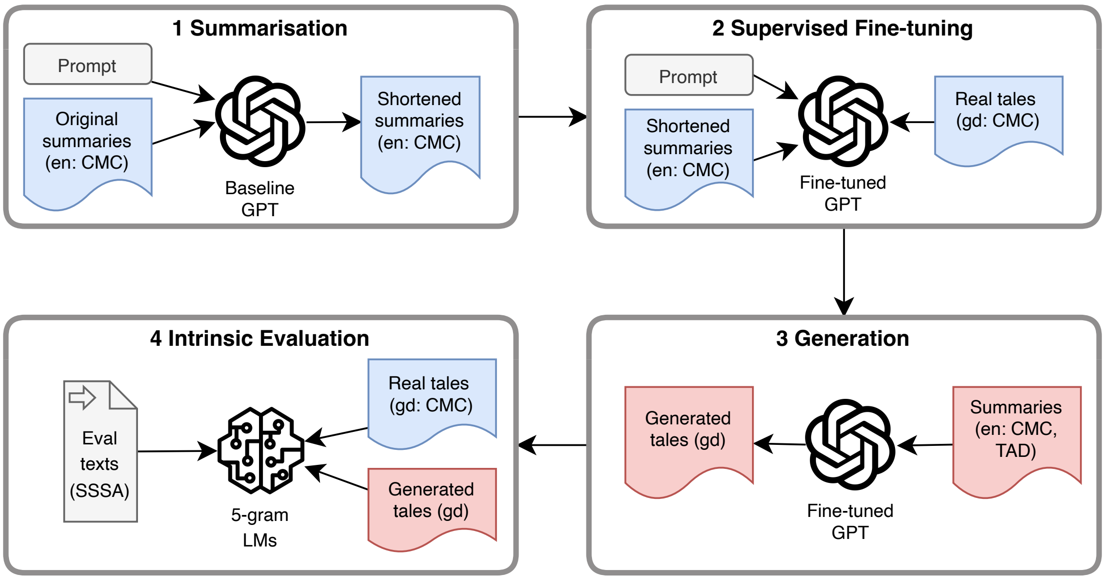

Figure 1: Training and evaluation pipeline. The summarisation prompt was ‘Summarise the given text in 6 to 7 sentences’. The supervised fine-tuning and generation prompt was ‘You will receive a summary in English. Expand the summary into a longer traditional narrative in Scottish Gaelic’. The datasets in Table 1 map to the pipeline’s steps as: Tr (1, 2); Tr-100 (1, 2, 4); Val (1, 2); Gen (3); Eval (4). Language key: en ‘English’; gd ‘Gaelic’.

Qi et al., 2024). In general, they find that current models like GPT-4o can produce coherent narratives up to about 2,000 words in English, but that text quality degrades in step with length after this point (Que et al., 2024). It is worth noting that transcriptions of Gaelic traditional narrative often exceed this word count (see Table 2). Additionally, MLLMs have been shown to have quality issues when generating synthetic text in LRLs (Robinson et al., 2023; Lai et al., 2023; Nguyen et al., 2023). Therefore, we expect to see more pronounced quality degradation in Gaelic long-form, synthetic texts due to weaker representations in pre-training corpora. To our knowledge, however, the area of LLM-based long-form text generation for LRLs remains unstudied.

2.3 Cross-lingual Transfer with MLLMs

An interesting property of MLLMs is their ability to share information between languages, known as *cross-lingual transfer*. To date, cross-lingual summarisation research involving MLLMs has focused on simultaneous translation and summarisation (see Wang et al., 2022). Here, instead, we attempt to share the knowledge encoded in GPT-4o’s English-based representations with Gaelic by expanding a summary in the former to a full text in the latter.

The quality of cross-lingual transfer between two languages depends on the degree of alignment between their feature spaces (Schmidt et al., 2022). Given the linguistic distance between Gaelic and English, and the presumed sparse Gaelic data in GPT-4o’s pre-training corpus, one might expect our generation quality to be quite low. Yet, in theory, we also may be able to leverage GPT-4o’s representations of a related and better-resourced Celtic language for our task, Irish. Positive transfer between higher-resource and lower-resource languages comes partially from the overlap in shared word-piece tokens (Conneau et al., 2020; Maguersse et al., 2020). Although Irish and Gaelic have somewhat distinct orthographies and grammar, they share a large proportion of lemmas (e.g. Irish *ballaí* ‘walls’ → *balla* ‘wall’; Gaelic *ballachan* ‘walls’ → *balla* ‘wall’). The implicit alignment between Irish and Gaelic sub-words, therefore, may benefit our task.

3 Datasets

To increase the available training data for Gaelic traditional narrative, we propose using XLTE, fine-tuning GPT-4o to produce long Gaelic narratives when prompted with an English summary of a traditional tale. Table 1 lists the six datasets used for this study and Figure 1 summarises the processing pipeline at a glance (see §4 for further details).

The ‘CMC’ data came from an orthographically standardised subset of the Calum Maclean Collection, an online corpus of Gaelic folktales.[5] The data contain transcriptions of Gaelic folktales paired with manually-produced English summaries and were split into training (‘Tr’), validation (‘Val’) and generation (‘Gen’) sets using a 80:10:10 ratio. The Tr and Val sets were used during supervised fine-tuning (Step 2 in Fig 1). The English summaries of the CMC Gen and the TAD Gen sets provided generation stimuli (Step 3 in Fig 1). We also created a CMC training set of the 100 longest tales (‘Tr-100’), to explore whether models fine-tuned on this set would produce longer, higher-quality outputs. As can be seen in Table 2, the word count distributions of the Tr sets are right-skewed; the median is a better measure of central tendency than the mean here.

The evaluation set (‘Eval’) comprised 158 Gaelic oral narrative texts from the Tale Archive of the School of Scottish Studies Archives (SSSA: see Sinclair et al., 2022). We used this as a reference set for computing the perplexity of various downstream n-gram LMs (see §4.4). Finally, a test set of 1,857 English summaries (‘TAD’) was used to prompt our best-performing fine-tuned model and assess XLTE for this use-case at scale. The TAD dataset comprised manual summaries of Gaelic folktales produced for the Tobar an Dualchais / Kist o Riches project[6] and is orthogonal to the CMC Gen set.

Source	Lang	Set	N	Words-gd
CMC	en/gd	Tr	384	276,958
CMC	en/gd	Tr-100	100	203,447
CMC	en/gd	Val	48	33,989
CMC	en	Gen	49	91,982
SSSA	gd	Eval	158	729,867
TAD	en	Gen	1,857	N/A

Table 1: Dataset statistics (gd: Gaelic; en: English)

4 Methodology

The processing pipeline, visualised in Figure 1, consists of four main steps. In Step 1 (‘Summarisation’), we prompt the baseline GPT-4o model to condense the human-generated English summary for each narrative in the training set. This step ensures consistency in summary length between the training and generation sets (see Table 1) and creates the paired data necessary for supervised fine-tuning (SFT). Step 2 (‘Supervised Fine-Tuning’) involves adapting the GPT-4o model to produce a naturalistic Gaelic narrative when given a plausible English summary. In Step 3 (‘Generation’), we synthesise a corpus of Gaelic narratives using the fine-tuned model, prompting it with authentic English summaries from held-out generation sets. Finally, in Step 4 (‘Intrinsic Evaluation’), we construct an n-gram language model (LM) from both the generated and genuine texts and evaluate their predictive accuracy on a held-out evaluation set. The following subsections provide further details on this pipeline.

Dataset	Mean	Median	St Dev
Tr	807.9	336	1747.6
Tr-100	2811.9	1652	3099.6

Table 2: Word count statistics of CMC training sets

4.1 Summarisation

While the TAD English summaries were 6.4 sentences and 78.8 words long on average, the original CMC English summaries were 14.2 sentences and 268.7 words long. We prompted the baseline GPT-4o model to condense the CMC summaries and equalise them with the TAD summaries’ average length. To accomplish this, we used the *system message* (‘prompt’), ‘Summarise the given text in 6 to 7 sentences’, and the following hyperparameter settings: n = 1, temperature = 1; top-p = 0.85; presence-penalty = 0.2; frequency-penalty = 0 and max-tokens = 250.

4.2 Supervised Fine-tuning

We assessed whether SFT would benefit this use-case or if the baseline GPT-4o model was sufficient. The SFT prompt was the same as that used for generation (Step 3): ‘You will receive a summary in English. Expand the summary into a longer traditional narrative in Scottish Gaelic’. After experimentation with hyperparameters, we fine-tuned for 3 epochs with a batch size of 1 and a learning rate multiplier of 2. As is standard, we monitored loss on both the training set and validation set to mitigate the risk of over-fitting.

---

[5] `https://www.calum-maclean-project.celtscot.ed.ac.uk/home/`

[6] `https://www.tobarandualchais.co.uk`. Note that a word count is not listed for the TAD source in Table 1 given that it consists of English summaries only; it contains no Gaelic text.

4.3 Generation

During generation, we prompted the models to expand the English summaries to longer Gaelic texts using the same system message as in the SFT step (see §4.2). After initial testing, we determined that the following hyperparameters achieve useful textual diversity, attenuate repetition and produce longer outputs: n = 1, temperature = 1; top-p = 0.5; presence-penalty = 0.3; frequency-penalty = 0.2 and max-tokens = 1000.

4.4 Evaluation Procedures

We deploy the following intrinsic evaluation metrics to assess the quality and performance of the generated texts:

- *mean word count* (MWC), which measures a GPT model's productivity; all things being equal, higher is preferred, due to lower generation costs and processing time
- *mean English to Gaelic ratio* (en:gd), a measure of code-switching levels, where lower is generally better;[7]
- *neologisms per total word count* (Neo), an estimation of hallucinated and nonce words, where lower is better;
- *perplexity* (PPL), which indicates the extent to which a LM predicts a textual input, where lower is better.

The English to Gaelic ratio is computed by calculating how many tokens in a given text occur in a large English dictionary, divided by how many occur in a large Gaelic dictionary. Tokens found in neither dictionary are considered neologisms[8] – hallucinated and otherwise out-of-dictionary tokens, many of which appear to be a by-product of Byte Pair Encoding (BPE) (Iwamoto and Kanayama, 2024). Perplexity measures the predictive accuracy of a LM against a reference text. Mathematically, it is the inverse of the geometric mean of the probabilities that a LM assigns to a text (Brown et al., 1992). Although perplexity is a commonly-deployed proxy for 'output quality', we acknowledge that it has an uneven relationship with human annotator scores (Stureborg et al., 2024). Additionally, it can be affected by superficial features such as text length (Meister and Cotterell, 2021) and punctuation (Wang et al., 2023).

To calculate perplexity on our texts, we lowercase and normalise them and then train a BPE tokeniser (Gage, 1994; Sennrich et al., 2016) on the full Gaelic narrative corpus, tokenising all input texts. Next, we train a 5-gram LM for the generated text with modified Kneser-Ney smoothing and no pruning, using the KenLM package (Heafield et al., 2013). We also train a LM using genuine narratives from the Tr-100 training set for comparison. As mentioned in §3, perplexity is calculated against the Eval set, comprising the SSSA texts.

5 Results and Discussion

In this section, we provide and discuss our experimental results. To recapitulate, our aim is to use XLTE to generate a synthetic corpus of Gaelic traditional narrative that can be used for downstream NLP tasks, such as training an external LM for an ASR system. Although this application is outwith the present scope, it will be examined in future research.[9]

5.1 Intrinsic Evaluation

We fine-tuned GPT-4o on the Tr (n=384) and Tr-100 (n=100) training sets (see Table 1) using the hyperparameter settings described in §4.2. We then trained n-gram language models on 48 examples of real training data (Train), data generated with the baseline GPT-4o model (4o-base) and the fine-tuned models (FT-100, FT-384). As shown in Table 3, the LMs associated with the fine-tuned models showed reduced perplexity scores on the Eval set, which indicates better predictive accuracy. Specifically, the LM associated with FT-100 achieved a PPL of 258.4, and FT-384 achieved 274.2, compared to the much higher PPL of 604.3 for the LM built from the baseline GPT-4o model's output. We observe that the FT-100 LM, which was trained upon the 100 longest narratives in the training set, achieved a slightly lower PPL value than the model trained on the full training set (FT-384), despite its smaller size. We expect that this occurred given that the FT-100 model generated longer texts on

---

[7] One could argue that it should resemble that of genuine narratives.

[8] We manually annotated a random sample of 100 automatically identified 'neologisms' and found 68% nonce words and compounds, 13% well-formed Gaelic compounds, 8% misspelled Gaelic words, 6% plausible dialectal variants, 3% Gaelic names and 2% English names.

[9] ASR for traditional narrative almost certainly will benefit from a language model built from a large, diverse dataset. Realistically, synthetic text from the target domain – if useful – would comprise only part of it.

average than the FT-384 model and had a slightly reduced neologism ratio.

Overall, the fine-tuned models are more productive than the baseline model, and have fewer English words and neologisms. These results suggest that fine-tuning, even on relatively small datasets, offers improvements in language modelling for this low-resource context. At the same time – and unsurprisingly – a LM built using the fine-tuning training data itself (i.e. the genuine transcriptions of Gaelic narrative in CMC) achieves a lower PPL still.[10] Thus, a model built directly on real data is likely to have a stronger grasp of the linguistic structures and discourse patterns inherent in the original texts.

Model	**PPL**	**MWC**	**en:gd**	**Neo**
FT-100	**258.4**	361.4	0.003	0.013
FT-384	274.2	330.5	0.003	0.014
4o-base	604.3	284.0	0.005	0.023
Train	140.2	630.3	0.007	0.007

Table 3: Intrinsic evaluation metrics for 48 generated (FT-100, FT-384 and 4o-base) and real narratives (Train). The results show improved performance of fine-tuned models over GPT-4o baseline on perplexity (PPL), mean word count (MWC), English to Gaelic ratio (en:gd) and neologism to total word count ratio (Neo). Metrics for the training set (Train) are provided for comparison.

To estimate the downstream benefits from using a larger amount of synthetic text, we use the FT-100 model to generate 1857 narratives (487,943 words) from the Gen set summaries (see §3) and concatenate these outputs with the original Tr-100 narrative texts. The results (see Table 4) indicate that scaling up the generated data improves the predictive power of the LMs built from them. Specifically, the PPL associated with a LM built using outputs derived from the Gen set summaries is lower (150.3) than that built using 48 outputs generated from the Tr-100 set summaries (258.4: see Table 3). Nevertheless, the LM built from authentic text achieves a superior PPL of 93.9, outperforming those built from generated data alone (PPL = 150.3) or concatenated real and generated data (PPL = 95.7). In comparison, we find that concatenating real and generated data enhances language models for the Gaelic news script domain (to be detailed in a future paper). To achieve a similar result in the narrative domain, we may need to increase the diversity of the SFT training data.[11]

Model	**PPL**	**MWC**	**en:gd**	**Neo**
Tr-100	**93.9**	1966.2	0.006	0.008
Gen	150.3	256.4	0.003	0.010
Concat	95.7	343.9	0.004	0.012

Table 4: Intrinsic evaluation metrics for the FT-100 training set (Tr-100, n=100), a large set of narratives generated with the FT-100 model (Gen, n=1857) and concatenated real and generated data (Concat, n=1957).

In sum, for this language, this MLLM and this use-case, one can generate a large synthetic training corpus with minimal effort and resources. Moreover, conducting XLTE with a fine-tuned GPT-4o model offers a clear performance boost over using the base model. The quality of the synthesised corpus, however, does not match human-produced data (cf. Alcoba Inciarte et al., 2024). This accords with research showing that MLLMs struggle to generate high quality output for LRLs (Robinson et al., 2023; Lai et al., 2023; Nguyen et al., 2023). Improvements may come from fine-tuning with larger training sets, further training epochs or increased learning rates. Nonetheless, it is possible that GPT-4o is under-resourced for Gaelic traditional narrative. Greater utility may come from training a LM on concatenated authentic and synthesised data when targeting domains that are better represented in GPT-4o's training data.

5.2 Stylistic Differences

To better understand divergences between the synthetic and genuine data for this use-case, we compared them for features that have been identified in previous research as being characteristic of Gaelic narrative, and spontaneous speech more broadly (Lamb, 2008). These are: 1) opening and closing formulas; 2) the narrative past-tense verb *arsa* 'quoth', used for direct quotation and 3) cosubordination. These features are likely amongst those implicit in perplexity differences although they are difficult to isolate programmatically. Here, we provide preliminary notes on the distributional and qualitative differences of these features in the datasets. We also briefly compare one synthetic narrative with its genuine counterpart.

---

[10] The longer average word count of the genuine narratives (630.3) likely accounts for some of the lower perplexity value; perplexity tends to negatively correlate with text length (Meister and Cotterell, 2021; Wang et al., 2023).

[11] As one reviewer points out, the better results for the news domain may come from the Gaelic pre-training corpus comprising mostly written text versus transcriptions of speech.

5.2.1 Opening and closing formulas

Traditional Gaelic storytellers often employ *formulas* in tales – stock phrases with myriad functions. The formula *bha siud ann (roimhe)* 'that was there (before)' is a common opening phrase and broadly equivalent to 'once upon a time' in English. While 17% of genuine tales from the Tr-100 set evince this formula, it occurs in none of the 1,857 generated tales from the Gen set. On the other hand, 18% of the generated tales begin with a more general variant, *bha* [noun phrase] *ann*, such as *bha tighearna ann roimhe* 'there was once a laird'. In comparison, only 5% of genuine tales have this variant. We believe that this finding can be explained by the fact that the phrase *bha* [NP] *ann* is generic and well-represented in GPT-4o's Gaelic pre-training corpus; it occurs in the language in other contexts. On the other hand, the prevalence of the more narrative-specific opening in the FT-100 training data was not high enough to induce the fine-tuned model to generate it.

Similarly, a common closing formula of Gaelic narratives is *agus dhealaich mi riutha* 'and I departed from them', which functions like the English phrase 'and they lived happily for ever after'. While this occurs in 26% of genuine tales, it does not appear in a single generated tale.

Taken together, these findings suggest that our fine-tuning conditions were not sufficient for GPT-4o to learn specific opening and closing formulas for Gaelic narrative. Although the model did learn a more general opening formula, it deployed it more frequently than would be expected in a genuine corpus. The complete lack of closing formulas may be a sign of attention decay (Li et al., 2024) during the fine-tuning process. Here, the model attends more to tokens or words that appear earlier in a sequence and less on tokens that appear later. Autoregressive LLMs generate text by iteratively predicting each token based on a probability distribution conditioned upon the input prompt and all previously-generated tokens. Thus, the model's prediction at each step relies on the context formed by the tokens generated so far. With increased context sizes, such as posed by longer narratives, performance can suffer. During generation, the lack of closing formulas may also be a sign of excessive weight on the local context at the detriment of the global context, a known problem with GPT models (Zhang et al., 2023). Based upon these results, it may be fruitful in the future to investigate diverging text quality and coherence in the heads and tails of long outputs.

5.2.2 Narrative direct-quotation verb

The defective verb *arsa/ars* 'quoth, said', or its variant *orsa/ors*, is a common feature of the storytelling register, where it is used to report direct speech. It often occurs in the register preferentially to another, more generic verb *thuirt* 'said', particularly in long stretches of dialogue. While the frequency for *arsa* is 37.6 per 1k words in the genuine texts, it is less than half that (16.4) in the synthetic texts. It is possible that an enhanced fine-tuning regimen, as discussed in §5.1, might encourage the model to adapt more closely to this and other stylistic features of the register.

5.2.3 Cosubordination

Cosubordination is a linguistic construction found in Gaelic and certain other languages (Van Valin and LaPolla, 1997) whereby a finite independent clause is coordinated with a non-finite dependent clause: e.g. *rinn e e agus iad nan cadal* 'He did it while [lit. and] they were asleep'. Cosubordination has been found to be closely associated with narrative registers (Lamb, 2008).

Using regular expressions, we searched for a type of cosubordination that is signposted by the occurrence of *agus* 'and' and a pronoun. In genuine narrative texts, this construction appeared 0.8 times per 1k words, but in synthetic texts, it appeared at 1/4 this rate: only 0.2 times per 1k words. Again, by strengthening the fine-tuning procedure – through increased stimuli or hyperparameter modification, or both – it may be possible to increase its appearance in synthetic texts. At the same time, if cosubordination and other constructions typical of traditional narrative are uncommon in pre-training data, it may be impossible to enhance the idiomaticity of synthetic text for this use-case beyond a certain margin. Of course, this may or not be detrimental to a particular application. For instance, if the synthetic text is used to train an n-gram LM, positive impacts on downstream tasks (e.g. ASR) may be possible even if the data are not fully concordant with the target domain.

5.2.4 Comparison of a synthetic and genuine narrative text

To further probe the fine-tuned model's output, we compared a synthetic text generated from one summary stimulus with a transcription of the original

Gaelic audio. These texts are available in Appendix A.[12]

The clearest difference between the synthetic and authentic text is that the former is shorter: its word count is 172, in contrast to 384 for the genuine text. Overall, we found generated text lengths somewhat labile; word counts for ten repeated generations in OpenAI's playground – using this summary and the hyperparameters stated in §5.1 – ranged from 90 to 561, with a mean of 196.

To expand a summary, the model must deploy its ingested knowledge to fill in the gaps logically. It is therefore useful to consider what the model adds to the synthetic text beyond the information that the summary provides. In this synthetic text example, the key place-name has mutated from Gerinish (*Geirinis*) to Garrynomonie (*Gearraidh na Mònadh*), another township in South Uist. The model infers from the label 'spinsters' that the women are socially isolated or awkward. Although this is a somewhat mild example of bias, it underlines the well-established tendency of LLMs to perpetuate negative stereotypes against women and older individuals (Zhao et al., 2024; Kamruzzaman et al., 2024; O'Connor and Liu, 2024).

Hallucination is well-known problem of LLM generation, whereby outputs are erroneous, contradictory or cannot be fact-checked (Ji et al., 2023). As introduced in §4.4, a sub-type of hallucination that affects our synthetic texts is LLM neologism, when a LLM outputs a word that cannot be found in a language's accepted lexicon (Iwamoto and Kanayama, 2024). Two neologisms occur in the synthetic output presented in Appendix A: *maighstir-sgioblaid* [recte *sgiobair*], which appears for 'captain', and *shùisg*, which is used for 'spit out' or 'vomit'. The first case is a compound of real and nonce elements. The modifier *maighstir* 'master' is a genuine word. The head noun, *sgioblaid*, appears to graft the element *sgiob-*, which is used in the words *sgioba* 'team' and *sgiobair* 'skipper', with *-laid*, an opaque ending that occurs in several real words (e.g. *trioblaid* 'trouble'). Regarding the second case, bodily functions in Gaelic are normally verbalised periphrastically (Lamb, 2024), that is using a generic verb (e.g. *dèan* 'do') along with the salient noun (e.g. *smugaid* 'spittle' → *dèan smugaid* 'spit' [lit. *make spit*]). Several natural Gaelic nouns meaning 'spit' begin with *s-*, but few verbs do. Additionally, no words with this connotation, to our knowledge, end in *-ùisg*. Yet the neologism *shùisg* 'spat, vomit' is perfectly understandable and even onomatopoeic. Although byte pair encoding is the transparent culprit for the first example, the second is more difficult to explain.

Another surprising element in the synthetic output is the appearance of gender agreement between pronouns and inanimate nouns: *brot* ('broth, soup', a masculine noun) maps correctly to *e* (a masculine pronoun) and *feòil* ('meat', a feminine noun) maps to *i* (a feminine pronoun). This type of agreement is waning in the spoken language. Its presence here suggests that, during pre-training, GPT-4o implicitly induced nominal gender as part of the embedding process and that MLLMs can share gender representations between higher- and lower-resource languages (Gonen et al., 2022). Future research could examine whether the publicly-available digital corpus for Gaelic is sufficient for MLLMs to reliably embed nominal gender representations, or if other related (e.g. Irish) and unrelated (e.g. French) languages produce a shared subspace for grammatical gender.

Turning to the genuine narrative example in Appendix A, we observe greater contextualisation and linguistic fluency, as expected from authentic storytelling. Specifically, the women are named, a historical context is provided (e.g. it happened twenty-three years ago and WWII rationing was in effect) and the location is given more precisely (i.e. the ship sunk at the Strait of Eriskay). In terms of lexis, the English word 'soup' is used instead of the less common Gaelic word *brot* 'broth, soup'. Also, the more charged label *Pàpanach* 'papist' is deployed instead of the more neutral *Caitligeach* 'Catholic', which appears in the synthetic text. Although the connotations of *Pàpanach* are milder in Gaelic than those of 'papist' in English, it would be unlikely to appear in formal Gaelic discourse. Although some MLLMs now offer basic support for LRLs like Gaelic, identifying and mitigating bias in these languages presents unique challenges due to toxicity and sparsity in training data, and cultural nuances. While addressing biases in HRLs is an active research area (Ferrara, 2023), additional work is needed to gauge how well current debiasing methods work for LRLs and how well they preserve linguistic and cultural diversity.

---

[12] We also provide a text generated with the baseline GPT-4o model for transparency. The baseline model's output is longer than that of the fine-tuned model, but it resembles a 19th century written homily more than a modern, vernacular folktale.

6 Conclusions

This study introduces cross-lingual text expansion (XLTE) as a scalable, LLM-driven method for mitigating data sparsity in low-resource languages and domains. By fine-tuning GPT-4o to expand English summaries into Gaelic narratives, we generated a substantial synthetic corpus that shows promise for improving language models. Our results demonstrate that supervised fine-tuning enhances performance over baseline GPT models, resulting in substantial reductions in perplexity and neologism. One surprising element found in synthetic texts was the appearance of gender-marked pronominal reference, which is fading from spontaneous spoken Gaelic. Nevertheless, qualitative analysis revealed stylistic discrepancies between synthetic and authentic narratives, particularly in terms of diminished formulaic language, narrative-specific verbs and cosubordination.

Future research should aim to narrow these stylistic gaps by improving the fine-tuning processes, for example by using more diverse, domain-specific training data and conducting further hyperparameter optimisation. One useful avenue would be to examine the relationship between top-p and hallucination (cf. Massarelli et al., 2020). Another would be to examine the effects of different prompts, such as explicitly declaring the desired word count. Given the limitations of working with a proprietary model, adapting the study to an open MLLM such as Meta's LLaMA (Etxaniz et al., 2024) would produce valuable insights. It also would be interesting to assess whether XLTE is more successful for domains that are better represented in pre-training corpora, such as news reportage. To conclude, we expect that XLTE is applicable to other domains and low-resource languages and has the potential to advance language modelling capabilities and downstream language technologies across diverse use-cases.

Limitations

The key limitation of this work is that it deploys a closed, proprietary large language model, GPT-4o. Beyond the basic details in OpenAI, 2024, OpenAI have not published information on GPT-4o's architecture, training data or fine-tuning procedures. This hinders the transparency and replicability of our study. For instance, we are unable to detect or remedy potential biases in the Gaelic training data or evaluate the model directly.

Acknowledgments

This work has benefited from the support of the Scottish Government (Grant name: 'Ecosystem for Interactive Speech Technologies') and the Arts and Humanities Research Council and Irish Research Council (Grant name: 'Decoding Hidden Heritages'; Grant numbers AH/W001934/1 and IRC/W001934/1). Thank you to the National Folklore Collection (University College Dublin) for permission to use data from the Calum Maclean Project, and to Tobar an Dualchais / Kist o Riches for providing access to the English and Gaelic summary data. Many thanks to Cailean Gordon for checking the Gaelic transcription and to Colin Jarvis of OpenAI for providing training credits and model access. Finally, we are grateful to the three anonymous reviewers for their constructive comments and advice.

Statement on Ethics

Institutional ethical review for this research was instigated on 2 March 2023 and granted on 13 March 2023 by the Ethics Officer of the host institution. Although no substantial risks are associated with this work, the authors acknowledge the environmental impact associated with pre-training large language models and, to a lesser extent, fine-tuning and generating from them. As a proprietary model was used for this study, it is impossible to provide an accurate measure of the carbon emissions associated with it at this time.

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.

Anthropic. 2024. Claude 3 model card. Accessed: 2024-10-08.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for NLI-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1318–1326.

Peter Brown, Della Pietra, Vincent J, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A tale of tails: Model collapse as a change of scaling laws. *Preprint*, arXiv:2402.07043.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Comput. Surv.*, 55(8).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. *Preprint*, arXiv:2403.20266.

Lucy Evans, William Lamb, Mark Sinclair, and Beatrice Alex. 2022. Developing automatic speech recognition for Scottish Gaelic. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 110–120, Marseille, France. European Language Resources Association.

Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. 2024. Can GPT-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, 31(10):2284–2293.

Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. 2022. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 67–77, Dublin, Ireland. Association for Computational Linguistics.

Xu Guo and Yiqiang Chen. 2024. Generative AI for synthetic data generation: Methods, challenges and the future. *Preprint*, arXiv:2403.04190.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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.

Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. CantonMT: Cantonese to English NMT platform with fine-tuned models using synthetic back-translation data. *Preprint*, arXiv:2403.11346.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal LLM. *TechRxiv*.

Ran Iwamoto and Hiroshi Kanayama. 2024. LLM neologism: Emergence of mutated characters due to byte encoding. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mahammed Kamruzzaman, Md. Minul Islam Shovon, and Gene Louis Kim. 2024. Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models. *Preprint*, arXiv:2309.08902.

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. 2008. *Scottish Gaelic Speech and Writing: Register Variation in an Endangered Language*. Cló Ollscoil na Banríona, Belfast.

William Lamb. 2024. *Scottish Gaelic: A Comprehensive Grammar*. Routledge Comprehensive Grammars. Routledge, Oxon.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling instruction (in)stability in language model dialogs. *Preprint*, arXiv:2402.10962.

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.

Hugh MacEachen. 1967. Muncaidh a theich bhon pholitician, agus dithis sheann mhaighdeannan. School of Scottish Studies Archives: SA1967.6.A6. Available from `https://www.tobarandualchais.co.uk/track/109839?l`.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Julie-Anne Meaney, Bea Alex, and William Lamb. 2024. Evaluating and adapting large language models to represent folktales in low-resource languages. In *The 4th International Conference on Natural Language Processing for Digital Humanities – NLP4DH 2024*.

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.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

National Records of Scotland. 2022. Scotland's census 2022. `https://www.scotlandscensus.gov.uk`. Accessed: 2024-08-14.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

OpenAI. 2024. GPT-4o system card. Accessed: 2024-10-04.

Sinead O'Connor and Helen Liu. 2024. Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI & SOCIETY*, 39(4):2045–2057.

Qianqian Qi, Lin Ni, Zhongsheng Wang, Libo Zhang, Jiamou Liu, and Michael Witbrock. 2024. Epic-level text generation with LLM through auto-prompted reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yiwei Qin, Graham Neubig, and Pengfei Liu. 2022. Searching for effective multilingual fine-tuning methods: A case study in summarization. *arXiv preprint arXiv:2212.05740*.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *Preprint*, arXiv:2409.16191.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Dusan Ilić-Sabo, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742.

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.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Mark Sinclair, William Lamb, and Beatrice Alex. 2022. Handwriting recognition for Scottish Gaelic. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 60–70.

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

Fahim Sufi. 2024. Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation. *Information*, 15(2):99.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *Preprint*, arXiv:2407.13248.

Robert D Van Valin and Randy J LaPolla. 1997. *Syntax: Structure, meaning, and function*. Cambridge University Press.

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.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2023. Perplexity from PLM is unreliable for evaluating text quality. *Preprint*, arXiv:2210.05892.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. *arXiv preprint arXiv:2402.17119*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *Preprint*, arXiv:2210.06774.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *Preprint*, arXiv:2403.00277.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Example of Synthetic and Genuine Narrative Text

System Message: You will receive a summary in English. Expand the summary into a longer traditional narrative in Scottish Gaelic.

User Message: The escaped monkey and the two old spinsters. A monkey escaped when the Politician sank and reached Gerinish. It went into the house of two old spinsters, who thought it was the captain. They gave it soup, which it spat out. It did the same with meat. It was a Friday, and one suggested it must be a Catholic, because they don't take meat on Fridays.

Synthetic Output from Fine-Tuned GPT-4o Model: *Nuair a chaidh am Politician fodha, theich muncaidh às. Chaidh e air tìr ann an Gearraidh na Mònadh. Bha dà chailleach thall an sin. Cha robh iad a' faicinn mòran dhaoine. Agus dh'fhalbh iad sìos gu taobh na tràghad a' coimhead dè bha a' tachairt agus chunnaic iad am muncaidh a' tighinn suas bhon tràigh. 'Ò,' thuirt tè dhiubh, 'seo fear de na mairbh a thàinig beò.' 'Ò, 's e,' thuirt an tè eile. 'Cò th' ann?' 'Tha mi cinnteach gur e am maighstir-sgioblaid a th' ann.' 'Ò, 's e,' thuirt i fhèin. 'Dè nì sinn ris?' 'Gabhamaid a-staigh e.'*

Thàinig am muncaidh suas chun an taighe agus dh'fhosgail iad an doras is ghabh iad a-staigh e. Chuir iad sìos aig an teine e. Thug iad brot dha agus shùisg e air feadh an ùrlair e. Thug iad feòil dha agus shùisg e air feadh an ùrlair i. Bha Dihaoine ann. Thuirt tè dhiubh: 'Tha mi cinnteach gur e Caitligeach a th' ann, agus nach eil e ag ithe feòla Dihaoine.'

Human Translation of Synthetic Output from Fine-Tuned GPT-4o Model: When the Politician sank, a monkey escaped. He went ashore in Garrynamonie. There were two old ladies over there. They did not see many people. And they went down to the beach to see what was happening and they saw this monkey coming up from the beach. 'Oh,' said one of them, 'this is one of the dead who survived.' 'Oh, yes,' said the other. 'Who is it?' 'I'm sure he's the captain.' 'Oh, it is,' she said herself. 'What shall we do with it?' 'Let's take him in.' The monkey came up to the house and they opened the door and took him inside. They put him down at the fire. They gave him soup and he spat it all over the floor. They gave him meat and he spat it all over the floor. It was Friday. One of them said: 'I'm sure he's a Catholic since he doesn't eat meat on Friday.'

Synthetic Output from Baseline GPT-4o Model: *Bha muncaidh air teicheadh nuair a chaidh am Poileataigs fodha agus ràinig e Geàrraidh na h-Aibhne. Chaidh e a-steach do thaigh dà sheann mhaighdean, a bha den bheachd gur e an caiptean a bh' ann. Thug iad brot dha, ach shèid e a-mach e. Rinn e an aon rud leis a' mhart-fheòil. Bha Dihaoine ann, agus mhol tè dhiubh gum feumadh e bhith na Chaitligeach, oir cha ghabh iad feòil air Dihaoine.*

A-nis, leig dhuinn an sgeulachd seo a leudachadh ann an Gàidhlig thraidiseanta.

Bha là fada is duilich aig muinntir Geàrraidh na h-Aibhne nuair a chaidh am bàta, ris an canar am Poileataigs, fodha ann an stoirm mhòr. Ach cha robh fios aig duine sam bith gun robh muncaidh air bòrd a' bhàta sin. Bha an t-ainmhidhean glic seo air teicheadh às an tubaist agus rinn e air tìr gu sàbhailte. An dèidh siubhal tro fhiadh-bheatha na sgìre, ràinig e mu dheireadh taigh beag air iomall a' bhaile far an robh dà sheann mhaighdean a' fuireach.

Nuair a chunnaic na seann mhaighdeanan am muncaidh a' tighinn a-steach don taigh aca, bha iad cinnteach gur e an caiptean bhon bhàta briste a bh' ann. "Thig a-steach, thig a-steach," thuirt tè dhiubh gu càirdeil, "feumaidh tu bhith sgìth is acrach." Dh'fhairich iad co-fhaireachdainn mòr ris agus chuir iad romhpa biadh blàth ullachadh dha.

Thug iad brot blasta dha, ach nuair a dh'fheuch am muncaidh ri ithe, shèid e a-mach e le tàmailt. Bha na seann mhaighdeanan fo iongnadh ach smaoinich iad gur dòcha nach robh e dèidheil air brot. Mar sin, chuir iad pìos mòr de mhart-fheòil air beulaibh an aoigh neo-àbhaisteach aca. Ach cha b' fhada gus an do rinn am muncaidh an aon rud ris a' mhart-fheòil—shèid e air falbh i gun mòran beachd.

Bha na seann mhaighdeanan fo chlisg agus thòisich iad air bruidhinn ri chèile mu dheidhinn nan tachartasan neònach seo. "Tha rudeigin gu math neònach mun duine seo," thuirt tè dhiubh. "Tha e Dihaoine an-diugh," thuirt am fear eile gu h-obann, "agus tha mi cinnteach gur e Caitligeach a th' ann. Chan eil Caitligich ag ithe feòil air Dihaoine."

Agus mar sin, dh'aontaich iad gum bu chòir dhaibh urram a thoirt do chreideamh am "caiptein" ùr aca agus thòisich iad air biadh freagarrach ullachadh dha airson Dihaoine—gun fheòil idir.

Mar sin thàinig deireadh air latha iongantach anns an robh dà sheann mhaighdean agus muncaidh teicheadh ag ionnsachadh mu dheidhinn cho cudromach sa tha urram agus tuigse eadar daoine (agus ainmhidhean) bho dhiofar chùl-raointean agus creideamhan.

Human Translation of Synthetic Output from Baseline GPT-4o Model: A monkey had escaped when the Politics sank and he arrived in Garynahine [NB: this is a township in Lewis, many islands away.]. He went into the house of two old maidens, who believed that he was the captain. They gave him soup, but he spat it out. He did the same thing with beef. It was Friday, and one of them suggested that he must be a Catholic, since they don't eat meat on Friday.

Now, allow us to expand this story in traditional Gaelic.

It was a long and difficult day for the people of Garynahine when the boat called Politics [*sic.*] sank in a big storm. But nobody knew that there was a monkey on board that vessel. This clever animals [*sic.*] had fled from the accident and he managed to get to land safely. After travelling through the wildlife [*sic.*] of the region, he finally arrived at a small house on the edge of the township where two old maidens were living.

When the old maidens saw the monkey coming into their house, they were sure that he was the captain of the ship. "Come in, come in," said one of them kindly, "you must be tired and hungry." They felt great sympathy for him and they set out to prepare some warm food for him.

They gave him some tasty soup, but when the monkey tried to eat it, he spat it out with shame. The two old maidens were confused but they figured that he wasn't keen on soup. Then they put a big piece of beef before their unusual guest. But it wasn't long until the monkey did the same thing with the beef—he spat it out without further consideration.

The two old maidens were startled and they began to discuss these strange events. "There is something very unusual about this fellow," one of them said. "It is Friday today," the other one [NB: masculine pronoun used] said suddenly, "and I'm sure that he is a Catholic. Catholics don't eat meat on Friday."

Then they agreed that they should respect the beliefs of their new "captain" and they began to prepare appropriate food for him for Friday–without any meat.

As such, the extraordinary day ended in which two old maidens and a monkey escaping [*sic.*] were learning about how important it is for respect and understanding to exist between people (and animals) from different backgrounds and belief-systems.

Human Transcription of Original Audio (MacEachen, 1967): *Uel bha siud ma-tà bho chionn trì bliadhna fichead air ais, cha chreid mi nach e a th' ann bhon a chaidh Am Politician air an sgeir ann an Caolas Èirisgeigh. Agus co-dhiù nuair a chaidh i air an sgeir, dh'fhàg an sgiobair is an criubha, dh'fhàg iad i uile gu lèir. Agus gu dè a bh' ac' air bòrd ach muncaidh. Agus theich am muncaidh 's rinneadh a thaighean cuideachd. Is bha e caran mu chuairt air feadh an eilein an sin – air ais 's air aghaidh – is cha robh e faighinn gu robh e dèanadh a dhachaigh an àite sam bith. Ach co-dhiù thàinig e dhan taigh a bha seo ann an Gèirinis. Agus bha dà sheann mhaighdeann ann, Ceit agus Mòr. Agus bha iad a' gabhail an dìnnear agus mar a tha fhios againn uile gu lèir bha coupons air a h-uile nì an àm a' chogaidh agus chan fhaigheadh tu ach beagan de dh'fheòil is beagan dhen a h-uile sìon. 'S ann le na coupons a bha thu ga fhaighinn, co-dhiù. Agus thàinig e ... bha iad a' gabhail an dìnnear, an dà sheann mhaighdeann a bha seo, agus dìreach cò thàinig a-staigh an dorast ach e seo, an giobal a bha seo, agus choimhead na boireannaich mu chuairt agus thuirt iad riutha fhèin, 'Ò an duine bochd. Sgiobair a' bhàta is chaidh i air an sgeir. Bheir sinn dha a dhìnnear.' Dh'èirich Mòr agus fhuair i soup dhan duine a thàinig a-staigh, dhan choigreach a thàinig a-staigh, agus bha . . . shuidh e aig a' bhòrd còmhla riutha. Fhuair e spàin 's dar a thòisich e air blasad air an soup, chuireadh e dhan bheul e is bheireadh e a-mach e 's spriodadh e air feadh an taigh e. Agus an sin, thuirt an dàrna tè ris an tè eile, 'Cha thoil leis soup,' thuirt i. 'Bheir sinn dha feòil 's buntàta.' Thug iad feòil 's buntàta dha. Thòisich cagnadh. Thilgeadh e pìos dheth an-dràsta air Mòr is pìos eile air Ceit agus ... nuair a bheireadh e treis air a' chagnadh. 'Ò an creutair, tha mise tuigsinn dè a th' ann,' thuirt i, 'ceart gu leòr. 'S e Dihaoine a th' ann an diugh! 'S e Pàpanach a th' ann is cha ghabh e brod na feòil an diugh,' thuirt i.*

Machine-assisted Translation of Original Audio: Well, it must've been about twenty-three years ago, I reckon, when the Politician ran aground on the skerry at the Strait of Eriskay. Anyway, when it hit the skerry, the skipper and the crew just up and abandoned it. And what did they leave behind but a monkey. The monkey ran off and made itself at home on the island, wandering here and there, not really able to settle anywhere. Eventually, it made its way to a house in Gerinish, where two old spinsters, Kate and Sarah, were sitting down for their dinner. Now, as we all know, during the war everything was rationed – you could only get a little bit of meat or anything else, and you needed coupons for everything. Anyway, as they were eating, in comes this ragamuffin through the door. The women turned and said to themselves, 'Oh, the poor man. It must be the skipper from the boat that went aground on the skerry. Let's give him something to eat.' So, Sarah got up and fetched a bowl of soup for the stranger, who sat down at the table with them. He took a spoonful, but as soon as it touched his mouth, he spat it right out, spraying it all over the house. One of the women looked at the other and said, 'He doesn't like soup. Let's give him some meat and potatoes.' So, they gave him meat and potatoes. He started chewing but, after a bit, he spat it out too – first on Sarah, then on Kate. ‘Oh, the poor creature,’ she said. ‘I understand now – it’s Friday! He’s Catholic and can’t have any meat today.’