

Qomhrá: A Bilingual Irish-English Large Language Model

Joseph McInerney
Trinity College Dublin
jmcinern@tcd.ie

Abstract

This paper introduces Qomhrá, a bilingual Irish-English large language model (LLM), developed under low-resource constraints presenting a complete pipeline spanning bilingual continued pre-training, instruction tuning, and alignment from human preferences. Newly accessible Irish corpora and English text are mixed and curated to improve Irish performance while preserving English ability. 6 closed-weight LLMs are judged for their Irish text generation by a native speaker, a learner and other LLMs. Google’s Gemini-2.5-Pro is ranked the highest and is subsequently used to synthesise instruction tuning and human preference datasets. Two datasets are contributed leveraging Gemini-2.5-Pro: a 30K Irish-English parallel instruction tuning dataset and a 1K human preference dataset, generating *accepted* and *rejected* responses that show near perfect alignment with a native Irish speaker. Qomhrá is comprehensively evaluated across benchmarks testing translation, gender understanding, topic identification and world knowledge with gains of up to 29% in Irish and 44% in English. Qomhrá also undergoes instruction tuning and demonstrates clear progress in instruction following, crucial for chatbot functionality.

1 Introduction

Large language models (LLMs) are trained on vast amounts of data and efficiently capture complex patterns leveraging the transformer architecture (Vaswani et al., 2017). These models have demonstrated remarkable adaptability, excelling in various downstream tasks such as summarization, translation, and information retrieval (Radford and Narasimhan, 2018). However, these advances are being made disproportionately in high-resource languages.

Irish illustrates this problem acutely, lagging significantly behind other European languages in

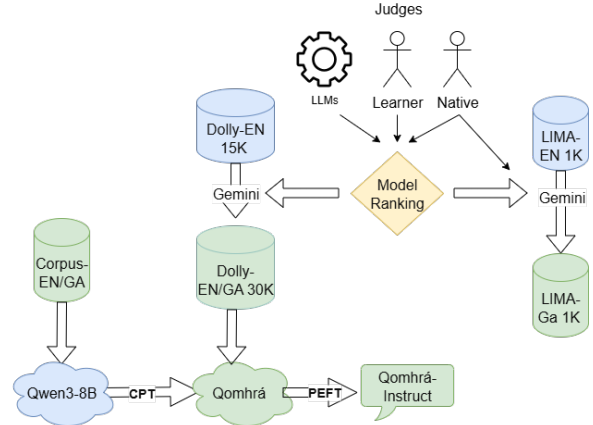


Figure 1: High-Level Pipeline Overview

terms of language technology despite its official status (Lynn, 2022). Previous efforts such as gaBERT (Barry et al., 2022) and UCCIX (Tran et al., 2024) demonstrate the potential of both encoder and decoder approaches for Irish. However, in order to develop a bilingual Irish-English chatbot, further LLM training stages of instruction tuning and preference alignment must be supported.

Therefore, this paper introduces **Qomhrá**, a bilingual Irish-English LLM. In developing Qomhrá, a framework for constructing instruction tuning and human preference datasets is outlined within low-resource language constraints. An overview can be seen in Fig 1.

The study is guided by three key questions.

1. Can bilingual continued pre-training (CPT) improve Irish performance without diminishing existing English capabilities?
2. Which closed-weight LLM provides the strongest Irish test generation for instruction dataset creation?
3. Can synthetic preference data generated by LLMs align with human judgments?

The scope of this work is bounded by computational hardware, Irish native-speaker access and data availability. Training is conducted at the 8B parameter scale without retraining tokenisers. Dataset creation is bound by the current standard of Irish observed in closed-weight LLMs. Evaluation is constrained by access to native Irish speakers for annotation relying on only one native speaker and one learner. These limitations reflect the necessary trade-offs when working with low-resource languages, while still enabling methodological insights and dataset contributions.

The contributions are threefold:

1. An Evaluation of bilingual CPT strategy.
2. A Comparative Evaluation of closed weight LLMs for Irish.
 - A 30K parallel English-Irish instruction tuning dataset adapted from the Dolly V2 (Conover et al., 2023) human-curated instruction tuning data set.
3. A novel method of human preference data synthesis requiring no labelled data in the low-resource language.
 - A 1K Irish human preference dataset adapted from the Less Is More for Alignment (LIMA) (Zhou et al., 2023) instruction tuning dataset.

The remainder of the paper is organised as follows. Section II reviews related work on LLM development for low-resource languages. Section III presents the methodology, including data preparation, model training and evaluation. Section IV reports experimental results, Section V discusses the implications of these results. Finally, Section VI concludes with key findings, limitations and future work.

2 Related Work

This section reviews the foundational and recent research in LLM development for low-resource languages. It describes the Irish language’s status as a low resource language and outlines a framework to develop an LLM relative to this context.

Foundationally, the transformer model architecture (Vaswani et al., 2017) enabled scalable sequence modelling via self-attention. This architecture permits LLMs like Qomhrá and others to process and learn complex relations from vast amounts

of data. LLM progress for Irish first involved a monolingual encoder gaBERT (Barry et al., 2022), which assembled Irish corpora, expanded model vocabulary and introduced Irish-specific benchmarks.

UCCIX (Tran et al., 2024) advanced this with a bilingual decoder LLM, once again leveraging large web corpora and tokeniser adaptation. Qomhrá builds on this further by outlining a full LLM chatbot pipeline including bilingual pre-training, instruction tuning and human feedback optimisation.

Definitions of *low-resource* vary from the socio-political, to human expertise and data availability (Nigatu et al., 2024). In the context of LLM development, Irish is classified as low-resource primarily due to the scarcity of data. More precisely, there is a lack of permissively licenced high-quality Irish language data, especially labelled data. This motivates the incorporation of newly available pre-training data and the creation of synthetic labelled data.

For low-resource languages, the adaptation of an existing LLM avoids training from scratch. This draws from the base model’s existing linguistic understanding to reduce training overhead. Various adaptation methods have been explored and their suitability in relation to developing Qomhrá is discussed.

Firstly, in-context learning (ICL) can adapt LLMs to a new domain (Cahyawijaya et al., 2024) but fails to produce models with intrinsic language competence and is more suitable as an inference-time aid.

Elsewhere, adaptation to new natural language domains has been achieved via supervised fine-tuning (Howard and Ruder, 2018). Parameter efficient fine-tuning (PEFT) (Rebuffi et al., 2017; Hu et al., 2021; Dettmers et al., 2023) has emerged as a computationally efficient method making it especially applicable in low-resource scenarios. However, a distinct new language requires a larger domain shift (Lu et al., 2025), better suited to CPT (Gururangan et al., 2020) if compute capacity allows.

CPT updates all parameters on unlabelled text and has proved effective for low-resource adaptation with the importance of incorporating English data in pre-training highlighted (Etzaniz et al., 2024; Tran et al., 2024). This mitigates a reduction in the model’s existing capabilities known as *Catastrophic Forgetting* (McCloskey and Cohen, 1989). UCCIX demonstrated that bitext data helps domain

transition, bridging the *stability gap* (Lange et al., 2023) but observed reduced performance in English language tasks. The Latxa, Basque model added English to the pre-training mixture, amounting to 17.75% but did not evaluate English language performance. Therefore, Qomhrá is pre-trained with English data and also evaluated on its English language performance.

Another challenge in adapting LLMs to low-resource languages is over-fragmentation in tokenisation. This makes training less efficient, shrinks context windows and slows token generation. exBERT (Tai et al., 2020) and others, showed heuristic-based methods (Yang et al., 2023) of expanding tokeniser vocabularies, even in extremely constrained scenarios (Yamaguchi et al., 2024). UCCIX successfully trained and merged an Irish-English tokeniser but the development of a bilingual tokeniser was out of scope for this project.

After adapting to the Irish language via CPT, the model must learn to follow instructions to become a chatbot. Instruction-tuned models generalise across tasks (Wei et al., 2022), but Irish lacks curated instruction data. Machine translation of existing English datasets has proved an effective strategy in creating high-quality instruction-tuning datasets (Laiyk et al., 2025; Etxaniz et al., 2024). It has even been effective at culturally aligning the LLM by seeding it with government/Wikipedia data (Laiyk et al., 2025). So in developing Qomhrá, a human-curated dataset is machine translated reflecting humanlike variance constrained by the quality of the translator.

Once instruction tuned, the final step of the LLM training pipeline is alignment from human feedback. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) established alignment by optimizing policy and reward models which is computationally complex. Direct Preference Optimisation (DPO) (Rafailov et al., 2023) simplifies the task to a pairwise classification. However both approaches require labelled data, where the inter-sample rating is known. Semi-supervised pipelines (Luo et al., 2025) use LLMs to generate synthetic candidates and pseudolabels for instruction tuning. Recent work with Kazakh show this approach extended to DPO (Kadyrbek et al., 2025), motivating the synthesis of a pairwise human preference dataset.

For Irish, there is no existing instruction-tuning data to seed the preference data. Therefore a novel approach is taken that leverages native alignment

to justify scaling human preference data. This is achieved by prompting an LLM to generate *accepted* and *rejected* responses requiring neither LLM nor human annotation.

3 Methods

3.1 CPT

Previous CPT pipelines have involved training with the open-weight Meta Llama series of models. However, the most recent series of models are restricted for fine-tuning in the European Union (Meta AI, 2025). The Qwen-3 model was selected due to its permissive Apache 2.0 licence, multilingual capabilities and state-of-the-art benchmark performance (Team, 2025). The 8B parameter dense model was selected as being the largest model possible given compute limitations.

3.1.1 English Text Collection

It was decided to collect data from the Irish parliament, called *the Dáil*. All debates of the last 10 years were collected using the Oireachtas API, amounting to 422M characters. The intention was to provide the model with culturally aligned information while also maintaining English performance. However, over-fitting to political themes was detected by testing the model’s text generation during training, illustrated in the supporting materials. Therefore the Dáil data was replaced by the Wikipedia dump 20220301 of 819M characters from the first 10K articles.

3.1.2 Irish Text Collection

The pre-training data from the development of the UCCIX Irish LLM was released open-source, this is comprised primarily of data crawled from the web and English-Irish bitext from the European Language Resource Coordination (ELRC). Additionally, a subset of the National Corpus of Irish was made available for this project under the CC BY-SA 4.0 licence. Due to copyright restrictions, the corpus was shuffled at sentence level and copyright protected data such as books was not shared with the author. An overview of the pre-training data and sources is shown in Table 1, where *En* represents English and *Ga* represents Irish.

3.2 Pre-processing

3.2.1 Deduplication

Near duplicate documents introduce redundancy in training. In the context of this project, the UCCIX data had already undergone pre-processing to

Table 1: Pre-training Corpora and Character Counts

| Source | Characters | Lang. | Prop. |
|-------------------|---------------|-------|--------|
| The Bible | 5M | Ga | 0.0017 |
| UCCIX_Leipzig | 13M | Ga | 0.0045 |
| UCCIX_ELRC | 17M | Ga-En | 0.0059 |
| UCCIX_Gawiki | 25M | Ga | 0.0087 |
| UCCIX_Gaparacrawl | 107M | Ga | 0.0372 |
| CNG | 549M | Ga | 0.1914 |
| UCCIX_Glot500 | 530M | Ga | 0.1848 |
| Wikipedia | 819M | En | 0.2851 |
| UCCIX_CulturaX | 1.2B | Ga | 0.4187 |
| Total | 2.869B | | 1.0000 |

remove duplicates. Therefore, the most important question was whether the CNG and Bible data were already *contained* (Broder, 1997) in the UCCIX data. Containment is defined below in 1, where $|A| < |B|$.

$$\frac{|L(A) \cap L(B)|}{|L(A)|} \quad (1)$$

To measure the containment the text was lower-cased and punctuation was removed. Documents were represented by sets of unique 5-gram deterministic hashes. The full inter-document containment can be viewed in the supporting materials, where the measured containment of CNG in the UCCIX was low. However, much of The Bible was contained in the UCCIX data but it was decided not to remove the Bible in the pre-training mixture due to its long-context, high quality and small size.

3.2.2 Segmentation

The end of document token `<|endof text|>` (Qwen Team, 2024) was inserted between samples across all data sources. For the Dáil, this meant between speaker utterances, for CNG, at the end of sentences and between each sample in the UCCIX data set. This is important for pre-training as it allows the model to distinguish clear context boundaries, leveraging a special token it is already familiar with. Each bitext sample was labelled [en] and [ga] to help the model identify between the two languages.

3.2.3 Vocabulary Expansion

The extension of the Qwen tokeniser was attempted for this project and proved effective with monolingual Irish tokenisation reducing tokenisation fragmentation in a test set by approximately 50%. However, the training of a bilingual Irish-English tokeniser was out of scope for this project. Instead, the multilingual base Qwen tokeniser was used to tokenise all text. Therefore Qomhrá’s training effi-

ciency and token generation speed is worse in Irish than in English.

3.2.4 Evaluation

To evaluate the pre-training, both the training loss and the evaluation loss were tracked to measure the optimisation to the new data. In order to quantify the model’s performance, it was tested with the same benchmarks as UCCIX. These benchmarks evaluate both Irish and English language skills across closed questions, translation tasks, and topic identification and Irish grammar. The model is compared with multiple other models.

The Llama 3.1-8B represents the baseline of an LLM of the same size not trained with a considerable amount of Irish data. This reflects the out-of-the-box performance of similarly sized open-weight LLMs not adapted to Irish. The UCCIX model enables comparison between two models having undergone CPT for Irish. English language benchmarks are of particular interest to measure the effectiveness of mixing a large proportion of English data in the pre-training data. Finally, different stages of Qomhrá’s development are compared, from the base Qwen3-8B-Base model to one epoch of CPT to the second and final epoch. This allows for insight into the effectiveness of CPT and also hyperparameter configuration.

3.2.5 Training Configuration and Hyper-parameters

For the training configuration, text was packed into 2048-token blocks significantly reducing the Qwen3-8B maximum context window of 128K due to memory constraints. Data was split 94/3/3 (train/validation/test) to monitor training stability. Bitext was prepended, in line with UCCIX methodology to smoothen domain shift, before mixing and deterministically shuffling the monolingual English and Irish data.

A per-device batch size of 1 with gradient accumulation ($\times 8$) was used to stay within memory constraints. DeepSpeed ZeRO-2 enabled data parallelism with both optimiser and gradient partitioning across GPUs. Gradient checkpointing reduced memory overhead. Training ran for 2 epochs to balance convergence with constraints, with validation monitoring from Weights & Biases and a test script to load model from checkpoint and run test generation every 3K steps to assess progress.

AdamW optimiser was used along with the default hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$,

Table 2: Models evaluated for Irish language generation.

| Provider | Flagship Model | Budget Model |
|-----------|-----------------|------------------|
| OpenAI | GPT-5 | GPT-5-mini |
| Anthropic | Claude-4-Sonnet | Claude-3.5-Haiku |
| Google | Gemini-2.5-Pro | Gemini-2.5-flash |

$\epsilon = 1e^{-8}$ and a learning rate of $1e^{-4}$ as shown to be effective in the training of BERT (Devlin et al., 2019).

The BF16 (Google Cloud, 2019) floating-point format was used as it approximates the dynamic range of FP32 with the 50% memory reduction FP16 provides (Kalamkar et al., 2019), which is important to maximize compute resources while maintaining training stability. The Hugging Face default linear warm up and scheduling is used without weight decay.

3.3 Instruction Tuning

In order to transition from next token prediction to chatbot functionality, Qomhrá is trained on instruction-response pairs. An existing English, human curated dataset Dolly V2 was selected to be translated to Irish. A translation of this dataset to Irish contributes a parallel Irish-English instruction-tuning dataset. The quality however, is bound by the machine translation model. As such, an experiment was set up to determine the strongest closed-source LLM for Irish.

Due to the scope of the project and requirements of the experiment, only 3 model providers were selected: OpenAI, Google and Anthropic. As this experiment was for text generation in a low resource scenario, the model costs were also taken into consideration. So, for each provider, both the most recent flagship model and a less expensive model were tested. These closed-source models are not evaluated specifically on their Irish language ability, so it was conceivable that a less expensive model could exhibit higher Irish performance than its flagship counterpart. The full list is shown in Table 2.

Each model was provided with Irish language text from the Irish parliament and Wikipedia data and asked to generate a prompt-response pair with reference to the text. This follows methodology applied to generate synthetic Kazakh instruction-tuning data, while also instilling culturally relevant knowledge (Laiyk et al., 2025). This was adopted as the task to evaluate the LLMs Irish language

generation abilities. The prompt used to generate instruction-response pairs was adapted from the Kazakh study.

The Kazakh study also showed moderate alignment between the LLM (GPT-4o) and expert human annotators. This is relevant in the low-resource scenario, as strong alignment allows for scaling because human annotators could be substituted with LLMs. This experiment aims to build on this idea of measuring annotation alignment between LLMs and a native Irish speaker to determine the viability of this strategy for Irish. To mitigate familiarity bias (Wataoka et al., 2024), all 3 LLM providers were used for evaluation. The *mode* of their collective responses was taken as their aggregate answer. The flagship model’s were chosen due to their superior performance.

Another annotator was tested, an Irish learner who self-identified as TEG B2 level. This follows on from the scaling logic, as there are many more learners of Irish than native speakers. This is reported by the Irish census where only 10% of Irish speakers indicated that they spoke Irish very well (Central Statistics Office, 2022).

After generating the prompt-response pairs, model comparisons were constructed. Model names were anonymised and ordering randomised. Comparisons were tagged with stable hash keys to keep track of progress and ensure human annotators received the same samples. Half of the samples were seeded with Wikipedia text and the other half Oireachtas. The annotator was presented with the initial source text and the instruction-response pair that model A and model B had generated, incorporating the text. They were asked the following question:

"Which Question–Answer pair exhibits a stronger command of Irish grammar and semantic coherence? Take the use of the reference text into account. If unsure, pick the one with a stronger display of Irish grammar. Choose A or B."

The question to annotate is the same for the LLMs except it clearly specified output formatting.

120 annotations were completed by both the learner and the native whereas the LLMs annotated 600 samples. So humans made a total of 8 comparisons per model pair and the LLMs made 20. The LLMs annotated more samples, demonstrating

their advantage of easy scalability. The order in which the models appeared was randomized and their identities anonymised to prevent bias.

The ranking model selected was the Bradley-Terry model. It was selected due to its strong performance dealing with small sample sizes and evenly distributed match-ups (Daynauth et al., 2025). It also provides stability as it does not require hyperparameter tuning unlike the Elo rating system. This makes it the ideal method to rank the head-to-head LLM comparisons.

Inter-annotator alignment was calculated using Cohen’s κ (Cohen, 1960), which measures alignment relative to random chance.

The strongest model, as determined by this experiment went on to translate the Dolly V2 instruction-tuning dataset. Qomhrá is then fine-tuned on the dataset using Low-Rank Adaptation (LoRA) to create Qomhrá-Instruct. Qomhrá-Instruct is then re-benchmarked on the same benchmarks as Qomhrá to evaluate the impact of the data set.

3.3.1 Training Configuration and Hyper-parameters

The Qwen chat template was applied to prompts in order for the Qomhrá base model to learn special tokens that indicate the flow of conversation. The thinking tags were removed as the goal was not to develop a reasoning model. Samples were padded to the maximum length with the pre-training document separator acting as the pad token, to enable packing for efficient training.

BF16 format was used and learning rate hyperparameters were tuned in line with the Unsloth LoRA Hyper-parameter Guide (Unsloth Documentation, 2025). The learning rate was $2e^{-4}$ with AdamW optimiser, rank 16, $\alpha = 32$, weight decay = 0.01 and 3% warm up ratio for fast adaptation.

The model was trained for 3 epochs balancing compute constraints and the improved performance associated with an increased number of epochs when instruction tuning (Lu et al., 2025).

3.4 Human Feedback

In order to evaluate the alignment of the human feedback dataset, another annotation experiment was set up. The LIMA dataset was selected to be translated due to its impressive results aligning LLMs with instruction following with only 1,000 examples. The top-ranked model from the previous LLM comparison experiment (Gemini-2.5-Pro)

was used to translate the data set.

Gemini-2.5-Pro was instructed to translate each sample from (prompt-En, response-En) to (prompt-Ga, response-1-Ga, response-2-Ga). Response-1 was to be a natural, direct and fluent translation whereas response-2 was to be unhelpful, non-idiomatic, inaccurate and awkward. The exact prompt is provided in the supporting materials. Given that the task was to evaluate whether the LLM was successful at generating *accepted* and *rejected* responses per human alignment, the only judge was the native Irish speaker.

An experiment was set up, where each annotation was a choice between response A and response B given the prompt. The ordering of the generated *accepted* and *rejected* responses was randomized and the original model-intended preference was stored with each annotation. This experiment determined whether the native agreed with the LLMs prescription of high-quality and low-quality to translations.

4 Experimentation and Results

4.1 Experimental Configuration

A comprehensive outline of software and hardware used is provided in the supporting materials for full reproducibility. Source code is available at <https://anonymous.4open.science/r/Qomhra-1E7D>.

4.2 Pre-training

Pre-training was carried out with 2 Nvidia H100 GPUs with 80GB VRAM for 34,360 steps with a total train-time of 44.196 hours.

The benchmarks are displayed in Table 3, the use of the same benchmarks as UCCIX allows for direct inter-model comparison across multiple tasks in both English and Irish. While more benchmarks would allow for more dimensions to be compared, this was out of scope for this project. The benchmarks are now described briefly.

The CLoze-Ga benchmark tests the model’s understanding of gender in Irish. SIB-Ga tests the model’s topic identification capability given a reference text. Irish Question Answer (IQA) is a question and answer benchmark developed by native speakers with topics relating to Ireland. The BLEU score measures translation accuracy from English to Irish and vice versa. Finally, the natural question (NQ) benchmark tests the models world knowledge with diverse closed questions. Detailed descriptions with examples can be found in the supporting

materials.

4.2.1 Irish Language and Culture

Firstly, Qomhrá outperforms the other 2 models in the Cloze benchmark closely followed by UCCIX. Its higher performance compared to Llama-3.1-8B is unsurprising as the gender information of vocabulary words can only be learned through exposure to the language. The highest IQA benchmarks in both English and Irish demonstrate a strong ability understand context contained in the question. This is useful for human-chatbot interactions where the chatbot needs to interpret reference text that the human provides in order to respond.

Ideally a future ablation study would be run keeping either the pre-training data or the base model fixed. This would allow the attribution of the benchmark improvements to the use of Qwen-3 or the addition of the CNG data.

4.2.2 Open-Ended Generation

Qomhrá’s poor performance on both BLEU benchmarks was of particular note as it had succeeded in other tasks. Upon inspecting the model’s responses, it was observed that the model frequently provided an accurate translation. However, the model did not stop generating tokens thereafter. These extra tokens are specifically factored in and penalised by the BLEU metric. The same is true for the NQs. This is a result of the model failing to learn to follow the task instructions when provided with few-shot examples.

4.2.3 Catastrophic Forgetting

Excluding the benchmarks where the model was evaluated under open-ended generation. The IQA-en benchmark was the only benchmark that evaluated the model’s English language capabilities. Qomhrá outperformed UCCIX by 29.21% in Irish as opposed to 44.35% points in English on the IQA benchmark. The information from these questions is the same so language must be the factor that caused Qomhrá to outperform UCCIX, indicating improved English performance retention for Qomhrá.

4.2.4 Training Epochs

Qomhrá did not improve performance with the second epoch of training. This is in line with expectations that base models saturate after 1 epoch of CPT (Lu et al., 2025). Therefore in the future, CPT would only trained for 1 epoch to prevent redundancy.

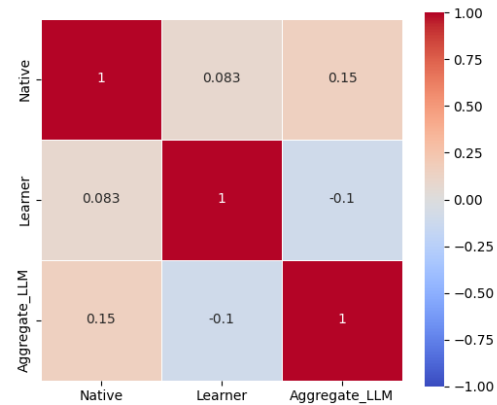


Figure 2: Inter-Annotator Agreement

4.3 Instruction Tuning

Table 4 shows that both the learner and the native were aligned in evaluating Gemini-2.5-Pro as the strongest model for Irish language text generation. Therefore, Gemini-2.5-Pro was used for subsequent synthetic dataset creation as the native speaker is considered the gold standard. In addition the native’s annotations were aligned with the costs of the models, ranking the 3 flagship models above the 3 budget models.

As shown in Fig 2, the annotators did not all agree. The agreement is roughly orthogonal indicating that agreement is almost equal to random chance. The inter-LLM annotation was also compared and there was weak alignment between all 3 models.

4.3.1 Benchmark Results

In Table 3, the significant improvement on the open-ended generation benchmarks BLEU (en2ga & ga2en) and NQ demonstrates the effectiveness of the instruction tuning dataset. A statistical analysis was conducted to evaluate the hypothesis that the improvement would be reflected in the open-ended response lengths.

Initial density plots were generated as displayed in the supporting materials, showing non-normal distribution across all benchmarks so the Mann-Whitney U test was selected. This determines whether the underlying distributions of word count lengths produced by the base model were *greater* than that of the instruct model.

As shown in Table 5, the base model (only CPT) has significantly longer responses than the instruct model across all benchmarks. For the NQ bench-

Table 3: Pre-training & Instruction-Tuning Benchmarking results

| Model | Cloze-Ga | SIB-Ga | IQA-Ga | IQA-En | BLEU en2ga | BLEU ga2en | NQ-En |
|-----------------|-------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Llama-3.1-8B | 0.59 | 0.7696 | 0.4861 | 0.7747 | 0.0880 | 0.4229 | 0.2767 |
| UCCIX | 0.75 | 0.7794 | 0.3889 | 0.3704 | 0.3334 | 0.4636 | 0.1668 |
| Qwen3-8B-Base | 0.44 | 0.6471 | 0.4633 | 0.8025 | 0.0154 | 0.2684 | 0.2590 |
| Qomhrá-1e-CPT | 0.85 | 0.8529 | 0.6810 | 0.8177 | 0.0368 | 0.0509 | 0.0374 |
| Qomhrá-2e-CPT | 0.86 | 0.8480 | 0.6810 | 0.8025 | 0.0363 | 0.0519 | 0.0355 |
| Qomhrá-Instruct | 0.88 | 0.8186 | 0.6760 | 0.7924 | <u>0.1167</u> | <u>0.0770</u> | <u>0.1269</u> |

Table 4: Bradley-Terry ranking of models. Abbreviations: GEM-Pro (Gemini-2.5-Pro), Claude-Sonnet (Claude-Sonnet-4), Claude-Haiku (Claude-3.5 Haiku), GEM-Flash (Gemini-2.5-flash).

| Rank | Native | Learner | LLM Agg. |
|------|---------------|---------------|---------------|
| 1 | GEM-Pro | GEM-Pro | GPT-5 |
| 2 | Claude-Sonnet | GPT-5-mini | GPT-5-mini |
| 3 | GPT-5 | Claude-Haiku | GEM-Pro |
| 4 | Claude-Haiku | GEM-Flash | Claude-Sonnet |
| 5 | GEM-Flash | GPT-5 | GEM-Flash |
| 6 | GPT-5-mini | Claude-Sonnet | Claude-Haiku |

mark, there is not 1 base model response that is shorter than any of the instruct model responses.

Table 5: Mann-Whitney U Test Results

| Benchmark | $Base_\mu$ | $Instruct_\mu$ | U_1 | p |
|-----------|------------|----------------|----------|----------------|
| en2ga | 127 | 54.2 | 224121 | $8.76e^{-105}$ |
| ga2en | 127.4 | 52.2 | 51457.5 | $1.74e^{-37}$ |
| NQ | 55.5 | 5.5 | 10526627 | 0 |

4.4 Human Preference Optimisation

Of the 91 preference annotations, the native speaker was aligned with Gemini-2.5-Pro for 90/91 examples, which gives a Cohen’s κ of 0.978, indicating near perfect alignment. This allows for the assumption that Gemini-2.5-Pro has created a dataset of prompt, response A and response B, where response A is preferred by humans. This allows the scaling to train with all 1,000 samples from the LIMA dataset.

5 Discussion

5.1 Pre-training

The success on both English and Irish benchmarks after CPT demonstrates the effectiveness of bilingual pre-training. This encourages the mixing of a significant amount of English data for others looking to develop bilingual LLMs without *catastrophic*

forgetting. This is limited by the breadth of benchmarking, where future work should aim to contribute new benchmarks evaluating dimensions like cultural alignment, helpfulness and summarisation.

5.2 Instruction Tuning

Claude, Gemini and Chat-GPT ranked a GPT model as the number one model for Irish language text generation. A plausible hypothesis for this is that these models have seen GPT distilled outputs and web content inducing bias in favour of GPT style outputs based on familiarity.

With moderate differences of inter-LLM agreement, it is difficult to discern a pattern, but this trend could indicate model specific biases which would be interesting to explore further. The poor native-learner and native-LLM agreement indicates that neither learners nor LLMs should be used to substitute for native speakers of Irish. However, the robustness of this generalisation is limited due to the small number of human participants. It is acknowledged that this diverges from other studies (Laiyk et al., 2025; Luo et al., 2025; Kadyrbek et al., 2025), where LLMs have proved useful in understanding human preferences.

Regarding open-ended benchmarking, the improved performance after instruction tuning was unsurprising as its intention is to improve unseen task performance (Wei et al., 2022) and motivates further instruction fine-tuning to become competitive.

5.3 Human Feedback

Unlike the instruction tuning, the LLM and the human showed near perfect alignment. This reflects the level of contrast between candidates in the tasks. The LLMs were unable to align with human preferences when the candidates were closer in quality in the instruction tuning experiment, whereas, when instructed, the LLM was able to generate high contrast samples that were in line with human preference.

This indicates a limitation of using LLMs for data synthesis in low-resource languages for scaling. Based on the results, it is suggested for LLMs to be used for tasks that require human preference alignment when the signal is clear. This should also be received with the consideration that there was only 1 native annotator.

6 Conclusion

This paper successfully pre-trained an open-weight LLM on Irish and English language data, improving on multiple benchmarks showing the effectiveness of mixing a high proportion of English data in the pre-training data.

Gemini-2.5-Pro was ranked as the strongest LLM for Irish language generation. This should guide others looking to integrate LLMs for Irish applications or research. In the future, more models could be tested with both human feedback and quantitative benchmarks for more comprehensive results.

Learners and LLMs were not aligned with the native speaker and thus, their use in Irish language annotation should be treated with caution, especially when the task is complex.

Unfortunately, a qualitative analysis of Irish produced between LLMs and machine translation models was out of scope. This is crucial future work to understand the optimal solution to synthesising Irish language data given the currently available models. This should also include qualitative analysis measuring hallucinations as this is an observed issue with LLMs for Irish translation (Castilho et al., 2025).

Gemini-2.5-Pro was almost aligned with the native speaker when tasked with human preference aligned data. This is encouraging as it demonstrates that previously costly human-preference datasets can be bootstrapped with no labelled data in the low-resource language. Training and benchmarking with this dataset would evaluate its effectiveness in action, and would require the development of new human alignment benchmarks for Irish.

Overall, this paper establishes key elements useful for the development of a chatbot for Irish. In doing so, it provides a framework for others to generate high-quality synthetic data.

6.1 Limitations

This paper is inherently limited by the number of languages explored and access to infrastructure,

data and human annotators. The bilingual pre-training strategy could lead to different results dependant on the relationship and distance between the two languages.

Due to computational constraints, an 8B parameter model was the maximum model size. Future work could evaluate whether performance scales further with model sized or whether the bottleneck is the data scarcity.

The lack of data for Irish makes the development of a robust model challenging. This motivates the collection and curation of high-quality permissively licenced Irish language text corpora. In parallel further data synthesis methods of Irish language data can be generated to supplement available data.

The lack of large scale access to native-speakers restricts the statistical robustness of findings of inter-annotator agreement and human-feedback alignment.

Ethical Considerations

The high proportion of pre-training data crawled from the web and lack of guard rails prevents the safeguarding of Qomhrá’s outputs, it should be used with this in mind. No personal information was stored from either annotator except for their Irish level ensuring data privacy. As mentioned, model identities in annotations were anonymised to prevent bias. The translation of English instruction data to Irish induces English-centric bias to the dataset and potential *translationese* not reflecting native spoken Irish.

Acknowledgments

I would like to thank my supervisor Dr. Barry Devereux who has supported and guided me throughout this project. I would also like to thank the ABAIR (abair.ie) research group. Finally, I would like to thank Jerry Sweeney and Khanh-Tung Tran of CloudCIX for the donation of computing resources.

References

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

- A.Z. Broder. 1997. [On the resemblance and containment of documents](#). In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29. IEEE.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. [LLMs are few-shot in-context low-resource language learners](#). 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 405–433, Mexico City, Mexico. Association for Computational Linguistics.
- Sheila Castilho, Zoe Fitzsimmons, Claire Holton, and Aoife Mc Donagh. 2025. [Synthetic fluency: Hallucinations, confabulations, and the creation of irish words in llm-generated translations](#). *CoRR*, abs/2504.07680.
- Central Statistics Office. 2022. Census of Population 2022 — Summary Results: Education and Irish Language. <https://www.cso.ie/en/releasesandpublications/ep/p-cpsr/censusofpopulation2022-summaryresults/educationandirishlanguage/>. Accessed: Aug. 27, 2025.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20:37 – 46.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. [Free dolly: Introducing the world’s first truly open instruction-tuned llm](#).
- Roland Daynauth, Christopher Clarke, Krisztian Flautner, Lingjia Tang, and Jason Mars. 2025. [Ranking unraveled: Recipes for LLM rankings in head-to-head AI combat](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26078–26091, Vienna, Austria. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NeurIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. [Latxa: An open language model and evaluation suite for Basque](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.
- Google Cloud. 2019. [Improve your model’s performance with bfloat16 | cloud tpu](#). Accessed: 2025-09-02.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Jeremy Howard and Sebastian Ruder. 2018. [Fine-tuned language models for text classification](#). *CoRR*, abs/1801.06146.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *CoRR*, abs/2106.09685.
- Nurgali Kadyrbek, Zhanseit Tuimebayev, Madina Mansurova, and Vítor Viegas. 2025. [The development of small-scale language models for low-resource languages, with a focus on kazakh and direct preference optimization](#). *Big Data and Cognitive Computing*, 9(5).
- Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. 2019. [A study of bfloat16 for deep learning training](#). *arXiv preprint*, arXiv:1905.12322.
- Nurkhan Laiyk, Daniil Orel, Rituraj Joshi, Maiya Goloburda, Yuxia Wang, Preslav Nakov, and Fajri Koto. 2025. [Instruction tuning on public government and cultural data for low-resource language: a case study in Kazakh](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14509–14538, Vienna, Austria. Association for Computational Linguistics.
- Matthias De Lange, Guido M van de Ven, and Tinne Tuytelaars. 2023. [Continual evaluation for lifelong learning: Identifying the stability gap](#). In *The*

- Eleventh International Conference on Learning Representations.*
- Wei Lu, Rachel K. Luu, and Markus J. Buehler. 2025. [Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities.](#) *npj Computational Materials*, 11:84.
- Junyu Luo, Xiao Luo, Xiushi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025. [Semi-supervised fine-tuning for large language models.](#) In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2795–2808, Albuquerque, New Mexico. Association for Computational Linguistics.
- Teresa Lynn. 2022. Report on the irish language. Deliverable D1.20 D1.20, European Language Equality (ELE).
- Michael McCloskey and Neal J. Cohen. 1989. [Catastrophic interference in connectionist networks: The sequential learning problem.](#) In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.
- Meta AI. 2025. [Llama 4 acceptable use policy.](#) Accessed: 2025-08-30.
- Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. [The zeno’s paradox of ‘low-resource’ languages.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.
- Qwen Team. 2024. [Key concepts - qwen documentation.](#) Accessed: Aug. 27, 2025.
- Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model.](#) In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. [Learning multiple visual domains with residual adapters.](#) *CoRR*, abs/1705.08045.
- Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. [exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3: Think deeper, act faster.](#) Accessed: 2025-08-30.
- Khanh-Tung Tran, Barry O’Sullivan, and Hoang D. Nguyen. 2024. [Uccix: Irish-excellence large language model.](#) In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, pages 4503–4506. IOS Press.
- Unsloth Documentation. 2025. Lora hyperparameters guide. <https://docs.unsloth.ai/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide>. Accessed: 2025-09-02.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS’17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *Safe Generative AI Workshop at NeurIPS 2024*.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners.](#) In *International Conference on Learning Representations*.
- Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. [How can we effectively expand the vocabulary of llms with 0.01gb of target language text?](#) Preprint, arXiv:2406.11477.
- Dong Yang, Xu Wang, and Remzi Celebi. 2023. [Expanding the vocabulary of bert for knowledge base construction.](#) In *Proceedings of the 1st Workshop on Knowledge Base Construction from Pre-trained Language Models (LM-KBC 2023) co-located with 22nd International Semantic Web Conference (ISWC 2023)*, volume 3577 of *CEUR Workshop Proceedings*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: Less is more for alignment.](#) In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.