Qomhrá: A Bilingual Irish and English Large Language Model

Joseph McInerney[1,3] mcinerjo@tcd.ie
Khanh-Tung Tran[2] 123128577@umail.ucc.ie
Liam Lonergan[1] llonerga@tcd.ie
Ailbhe Ní Chasaide[1] anichsid@tcd.ie
Neasa Ní Chiaráin[1] nichiarn@tcd.ie
Barry Devereux[3] b.devereux@qub.ac.uk

[1]Trinity College Dublin, [2]University College Cork, [3]Queen's University Belfast

Abstract

Large language model (LLM) research and development has overwhelmingly focused on the world's major languages, leading to under-representation of low-resource languages such as Irish. This paper introduces **Qomhrá**, a bilingual Irish and English LLM, developed under extremely low-resource constraints. A complete pipeline is outlined spanning bilingual continued pre-training, instruction tuning, and the synthesis of human preference data for future alignment training. We focus on the lack of scalable methods to create human preferences by proposing a novel method to synthesise such data by prompting an LLM to generate "accepted" and "rejected" responses, which we validate as aligning with L1 Irish speakers. To select an LLM for synthesis, we evaluate the top closed-weight LLMs for Irish language generation performance. Gemini-2.5-Pro is ranked highest by L1 and L2 Irish-speakers, diverging from LLM-as-a-Judge ratings, indicating a misalignment between current LLMs and the Irish-language community. Subsequently, we leverage Gemini-2.5-Pro to translate a large scale English-language instruction tuning dataset to Irish and to synthesise a first-of-its-kind Irish-language human preference dataset. We comprehensively evaluate Qomhrá across several benchmarks, testing translation, gender understanding, topic identification, and world knowledge; these evaluations show gains of up to 29% in Irish and 44% in English compared to the existing open-source Irish LLM baseline, UCCIX.

1 Introduction

Whilst progress has been made in speech synthesis and recognition for Irish (Lonergan et al., 2022), Irish remains the least-supported official European language in terms of language technology (Lynn, 2022). To address this, we present **Qomhrá**, an open-weights bilingual Irish and English LLM, developed under extremely low-resource constraints.

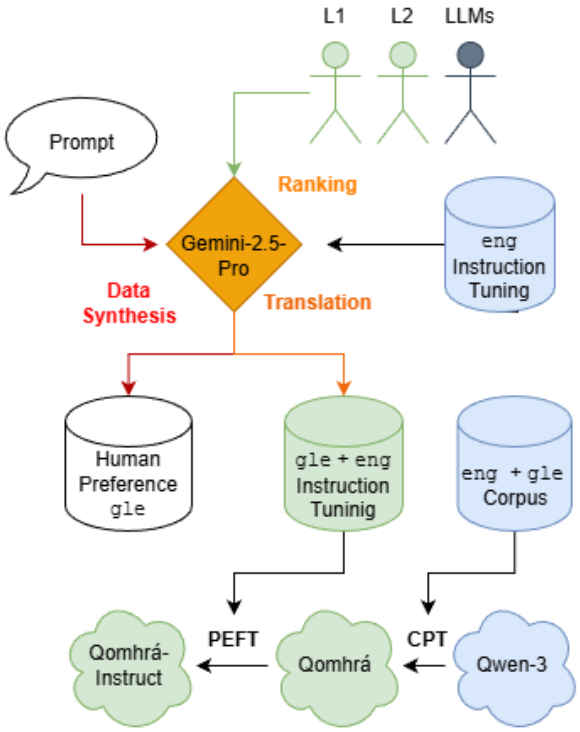

Figure 1: A High-Level Pipeline Overview of Data Synthesis, Model Ranking and LLM Training of the Qomhrá Irish-English language model.

Qomhrá supports the subsidiary development of chatbots and other NLP tools, applicable across education, translation, and public services, enabling access to language technologies for the Irish-language community. Crucially, Qomhrá fosters technological sovereignty for the Irish language by reducing reliance on proprietary API-based models. This mitigates risks associated with cost volatility and data privacy, while ensuring the community retains control over the adaptation of models to specific use cases.

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, we classify Irish as a low-resource lan-

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 189–199

guage due to data scarcity, or, more precisely, the lack of permissively-licensed labelled data for instruction tuning and alignment. This data requires explicit human annotation, making it particularly valuable for training and evaluating LLMs but also it is typically costly to create. Qomhrá is trained on over 3B characters of text which is multiple orders of magnitude less than foundation LLMs such as Llama-3 (Grattafiori et al., 2024).

To facilitate instruction tuning, we curate a labelled dataset by translating from English with a closed-weight model. To select this translator LLM, we evaluate budget and flagship models from top providers. We use three separate judges: an L1 speaker, an L2 speaker and the aggregate of three LLMs.

We include an L2 speaker to measure agreement with the L1 speaker, this informs a cost-quality trade-off, driven by the much higher proportion of L2 speakers. In the absence of specific statistics, we estimate the number of L1 speakers to be similar to the 71,968 people that speak Irish daily outside of education (Central Statistics Office, 2022). This is far fewer than the 1,873,997 that declared that they could speak Irish. This scarcity also motivates our small scale human annotation. Furthermore, we note that we do not control for dialect by only including an L1 speaker of Ulster Irish. Future work should measure the inter-annotator agreement across all three Irish dialects.

We further address labelled Irish data scarcity through the proposal of a novel synthesis framework that prompts the strongest available closed-weight model to translate an instruction-response pair from the high-resource language, English (eng), to the low-resource language, Irish (gle). The instruction remains constant but the LLM is prompted to generate a poor-quality and high-quality response mimicking human preference. This synthesis framework is applicable to other extreme low-resource language scenarios, where little is known about existing closed-weight LLM capabilities and annotated data is scarce. The framework is visualised in Fig 1 and investigates three key questions:

1. To what degree does bilingual continued pretraining (CPT) improve Irish performance without diminishing existing English capabilities?

2. Accounting for API cost, which closed-weight LLM exhibits optimal Irish language chatbot performance as judged by an L1 Irish speaker?

3. To what degree does synthetic preference data generated by SOTA LLMs align with human judgments?

The scope is bound by compute, data, and L1 speaker access. We train at the 8B scale without tokeniser retraining, and our evaluation relies on one L1 speaker and one L2 speaker. These restrictions are typical of low-resource research but still enable key methodological insights. **The contributions of our paper are five-fold:**

1. The release of Qomhrá, an 8B parameter open-weight Irish and English LLM for non-commercial purposes [1], including a language-aware quantized version [2].

2. A first-of-its-kind, human evaluation and ranking of the top available closed-weight LLMs for Irish.

3. A 30K sample parallel English-Irish instruction tuning dataset that significantly improves the LLM's ability to follow instructions [3].

4. A novel method to synthesise human preference data by prompting an LLM to generate "accepted" and "rejected" translations of an existing English prompt-response pair.

5. A 1K sample Irish human preference dataset that is shown to align with an L1 Irish speaker [4].

2 Related Work

2.1 LLMs for Low-Resource Languages

While major languages, particularly English, dominate interest in state-of-the-art LLM performance, there have also been efforts that have focused on adapting models to low-resource languages. Initial efforts for Irish focused on encoder-only models such as gaBERT and gaELECTRA (Barry et al., 2022). More recently, the focus has shifted to generative decoder models, such as UCCIX for Irish (Tran et al., 2024) and elsewhere, Latxa for Basque

[1] https://huggingface.co/jmcinern/Qomhra

[2] https://huggingface.co/jmcinern/Qomhra-AWQ

[3] https://huggingface.co/datasets/jmcinern/Dolly-V2-gle

[4] https://huggingface.co/datasets/jmcinern/LIMA-gle-DPO

(Etxaniz et al., 2024), adapting to their respective low-resource languages via continued pre-training (Gururangan et al., 2020). We note that both efforts report cases of catastrophic forgetting (McCloskey and Cohen, 1989), where English performance declines, and thus we include a significantly higher proportion of English training data for bilingual capabilities.

2.2 Instruction Tuning

For high-resource languages, instructions annotated by humans are abundant. In low-resource settings, they are often either low-quality or absent. A prevailing strategy is *translate-train* (Conneau et al., 2018), where data is translated from source to target language and then used to train the model. This is effective at augmenting low-resource instruction tuning data (Singh et al., 2024).

We adopt this methodology by translating the Dolly V2 dataset (Conover et al., 2023), but we extend the pipeline by evaluating the translator model first, since in low-resource settings model providers often do not evaluate low-resource languages. This extension helps others looking to select models for Irish data synthesis and highlights a key step for others seeking to synthesise data in other low-resource languages.

2.3 Alignment

After instruction following, the next step is to align models to human preferences, which has typically relied on Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), a procedure which is computationally complex and costly. Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers a more stable, data-efficient alternative but requires paired "chosen" and "rejected" response data, which does not exist for Irish. Synthetic data generation for DPO has been explored for other languages (Kadyrbek et al., 2025), often using a teacher model to score responses. We propose a simplified, novel approach that explicitly prompts a strong multilingual teacher (Gemini-2.5-Pro) to synthesise the preference signal itself by generating contrasting translation qualities, thereby bypassing the need for a fine-tuned teacher model.

3 Model Training Framework

3.1 Continued Pre-training Data

As with the GaBERT and UCCIX models, we use the available web-crawled Irish data, drawing from the de-duplicated dataset released by the UCCIX team (Tran et al., 2024). Additionally, a subset of the National Corpus of Irish was made available for this project under the CC BY-SA 4.0 license. Due to copyright restrictions, the corpus was shuffled at the sentence level and copyright-protected data was not shared. An overview of the pre-training data and sources is shown in Table 1. MinHash (5-grams) was used to detect duplicates.

Table 1: Pre-training corpora and character counts. Note: Total character count is 3.265B.

Source	Characters	Lang.	Prop.
The Bible	5M	gle	0.0015
UCCIX_Leipzig	13M	gle	0.0040
UCCIX_ELRC	17M	gle & eng	0.0052
UCCIX_Gawiki	25M	gle	0.0077
UCCIX_Gaparacrawl	107M	gle	0.0328
CNG	549M	gle	0.1681
UCCIX_Glot500	530M	gle	0.1623
Wikipedia	819M	eng	0.2508
UCCIX_CulturaX	1.2B	gle	0.3675
Total	**3.265B**		1.0000

For English text, we used the first 50K samples from Wikipedia dump 20220301 [5], to cover a diverse range of topics.

3.1.1 Segmentation

The end of document token `<|endoftext|>` (Qwen Team, 2024) was inserted between samples across all data sources, preventing the model from inferring dependencies between unrelated documents and leveraging a special token it is already trained with. For the Dáil dataset, this corresponded to between speaker utterances; for CNG, the end of text token was appended to each sentence (given the sentence-level data shuffling of this dataset); and for the UCCIX dataset between each textual sample in the dataset. Each parallel eng-gle sample was explicitly prepended with its respective language tag: "[eng]"/"[gle]".

3.1.2 Evaluation and Baselines

We quantify the model's performance with the same benchmarks as UCCIX. These benchmarks evaluate both Irish and English language skills across closed questions, translation tasks, topic identification, and Irish grammar. In order to measure the relative performance of Qomhrá, we compare against two models of the same size, one having undergone no CPT for Irish, Llama 3.1-8B, and

[5] https://dumps.wikimedia.org

the other having been adapted to Irish via CPT, UCCIX. To understand training dynamics, we evaluate the base model, the model after one epoch and the model after two epochs of CPT. This measures the impact of CPT on the model's performance and the extent to which further epochs improve adaption.

3.2 Continued Pre-training

For low-resource languages, the adaption of an existing LLM avoids training from scratch. This draws from the base model's existing linguistic understanding to reduce training overhead. We explore various adaptation methods and their suitability to developing Qomhrá.

While methods like in-context learning (Cahyawijaya et al., 2024) and parameter efficient fine-tuning (PEFT) (Hu et al., 2021; Dettmers et al., 2023) offer efficient adaptation, continued pre-training (CPT) (Gururangan et al., 2020) is better suited for the large domain shift of a new language (Lu et al., 2025).

3.2.1 Training Configuration and Hyper-parameters

For the CPT configuration, we pack the text into 2048-token blocks, significantly reducing the Qwen3-8B maximum context window of 128K due to memory constraints. We split the pre-training data 94:3:3 (train:validation:test) to monitor training stability. In line with UCCIX methodology, we prepend bitext to smoothen domain shift, before mixing and shuffling the monolingual English and Irish data. Shuffling is done deterministically for reproducibility.

A per-device batch size of one 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 two epochs to balance convergence with constraints, with validation monitoring from Weights & Biases (Biewald, 2020) and a test script to load model from checkpoint and run test generation every 3K steps to assess progress.

The AdamW optimiser was used along with the default hyper-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$ and a learning rate of $1e^{-4}$.

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 maximise compute resources while maintaining training stability.

3.3 Instruction Tuning

Instruction tuning transitions an LLM from a token predictor to a chatbot, useful for interfacing with humans. Qomhrá is trained on instruction-response pairs which requires significantly less data (Wei et al., 2022). We translate the Dolly V2 (Conover et al., 2023) dataset to Irish. A translation of this dataset contributes a parallel Irish and 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-weight LLM for Irish, outlined in Section 4.1.

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 dataset.

The Qwen chat template was applied to prompts. The thinking tags were removed as the goal was not to develop a reasoning model. 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.

3.4 Human Feedback

We propose a novel method that synthesises "accepted" and "rejected" responses, to allow for direct preference optimisation without any human annotation. This facilitates alignment in extremely low-resource settings, where access to L1 speakers is limited and costly. We instructed Gemini-2.5-Pro to generate paired translations for each English instruction: one high-quality ("accepted") and one low-quality ("rejected")

To do this, Gemini-2.5-Pro was instructed to translate each English instruction response sample, where the prompt is identical but the response varies, where an L1 speaker is expected to "accept" one response and "reject" the other. The LIMA dataset (Zhou et al., 2023) was selected for translation, due to its effectiveness at small scale. The top-ranked model, Gemini-2.5-Pro, was used to synthesise the preference dataset. The following is

the prompt we designed, specifying high contrast and strict JSON formatting for robust API parsing.

```
Translate the following
English Instruction and
response into Irish.
- response1 should be a
natural, direct and fluent
translation,
- response2 should be
a weak alternative, it
should be unhelpful, not
idiomatic, inaccurate,
awkward.
- The contrast in quality
of Irish should be very
high.
OUTPUT FORMAT (STRICT):
Return strict JSON with
exactly:
{
"instruction":
"<instruction in Irish>",
"response1":  "<much
better response in
Irish>",
"response2":  "<much worse
response in Irish>"
}
The following is the
English prompt-response
pair:
```

The English instruct-response pair was appended to the request.

A validation experiment was set up where the L1 speaker indicated preference between response A and response B. The in-pair ordering was randomly shuffled and the original model-intended preference was stored with each annotation. This experiment determined whether the L1 speaker agreed with the LLM's prescription of high-quality and low-quality translations.

4 Experimental Setup and Evaluation

4.1 Translator Selection Methodology

The top three commercial LLM providers were selected: OpenAI, Google, and Anthropic. For each provider, both the most recent flagship model and a less expensive model were tested. The budget model was considered as these model providers do not specifically evaluate Irish language capabilities and thus the flagship model is not necessarily the best or most cost-effective model. Therefore, we compare flagship vs budget performance which is particularly relevant for low-resource scenario data synthesis.

Table 2: Bradley-Terry ranking of models. Abbreviations: GEM-Pro (Gemini-2.5-Pro), Claude-Sonnet (Claude-4-Sonnet), Claude-Haiku (Claude-3.5 Haiku), GEM-Flash (Gemini-2.5-Flash).

Rank	L1 Speaker	L2 Speaker	LLMs
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

Each model was provided with Irish language text from the Irish parliament (Dáil) and Wikipedia and asked to generate a prompt-response pair. This follows methodology applied to generate synthetic Kazakh instruction tuning data (Laiyk et al., 2025).

120 annotations were completed by both the L2 and the L1 speaker, whereas the LLMs annotated all 600 samples. The ranking model selected was the Bradley-Terry model due to its strong performance dealing with small sample sizes (Daynauth et al., 2025).

4.2 Translator Selection Analysis

Table 2 shows that both the L2 and the L1 speaker 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 L1 speaker is considered the gold standard.

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-generated 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.

4.3 Pre-training

Pre-training was carried out with two 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.

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

Model	Cloze-gle	SIB-gle	IQA-gle	IQA-eng	BLEU eng2gle	BLEU gle2eng	NQ-eng
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	0.1167	0.0770	0.1269

Bold indicates the best performing model whereas underline indicates the best performing Qomhrá model. 1e and 2e refer to 1 and 2 epochs of CPT.

4.3.1 Benchmark Descriptions

To provide clarity on the evaluations, we provide a detailed description of the five benchmarks used as there are none currently available.

Cloze-gle (Tran et al., 2024) evaluates the model's grasp of Irish grammatical gender and person agreement (masculine/feminine/plural). The model is presented with three candidate sentences differing only by a single pronoun (e.g., *é, í, iad*). The log-likelihood (LL) of each full sentence is computed. Since the candidates differ by only one word, the sentence with the highest likelihood represents the model's prediction. In Table 4, the noun *faoiseamh* is masculine. The model assigns the highest likelihood (-38.2) to the sentence containing the masculine pronoun **é**, matching the target.

#	Candidate Sentence	LL	Target
1	Is iontach an faoiseamh **í** sin.	-43.4	No
2	Is iontach an faoiseamh **é** sin.	**-38.2**	**Yes**
3	Is iontach an faoiseamh **iad** sin.	-43.4	No

Table 4: Cloze-gle Example. The distinguishing pronoun is bolded.

SIB-gle (Adelani et al., 2024) tests topic modelling capabilities. Unlike the zero-shot Cloze task, this is a 10-shot evaluation. The model is presented with input text and must assign it to one of seven high-level topics (e.g., *eolaíocht/teicneolaíocht* [science/tech], *polaitíocht* [politics]). The model scores each topic label as a continuation; the label with the highest probability is selected. In Table 5, the text describes how a nuclear bomb is made and the model correctly assigns the highest probability to the science category. Only 2 of the 7 categories are displayed in Table 5 for brevity.

Text	*Oibríonn an chéad bhuama eamhnaithe ar an bprionsabal gur gá fuinneamh...*		
#	**Candidate Label**	**LL**	**Target**
1	eolaíocht/teicneolaíocht	**-0.017**	**Yes**
2	polaitíocht	-6.264	No

Table 5: SIB-gle Example (truncated).

IQA-gle & IQA-eng (Tran et al., 2024) assess multiple-choice question answering in both English and Irish using a 5-shot context. For a given question, the model is presented with four candidate answers. The log-likelihood of each candidate answer is calculated given the question context. The questions are related to Ireland and so this benchmark also measures cultural knowledge.

Table 6 illustrates an example of IQA-eng where the model correctly identifies that Irish is taught as a compulsory subject. The method is the same for IQA-gle.

Q	How is the Irish language taught in secondary education?		
#	**Candidate Answer**	**LL**	**Target**
1	As an optional subject	-1.7	No
2	As an extracurricular activity	-4.0	No
3	As a compulsory subject	**-0.3**	**Yes**
4	As a foreign language	-3.2	No

Table 6: IQA-eng Example.

BLEU (gle ↔ eng) (Lankford et al., 2022) evaluates translation quality via BLEU-4 in a 5-shot setting on health data. Their model trained on in-domain data serves as an upper bound with BLEU scores of 57.6 and 37.6 for gle2eng and eng2gle respectively. Table 7 shows a sample generation where Qomhrá's response omits the definitive plural article *na*, this difference is highlighted in red.

Source (eng)	Important notice: Latest information on Revenue services and tax and customs measures...
Reference (gle)	Fógra tábhachtach: An t-eolas is déanaí faoi sheirbhísí na gCoimisinéirí Ioncaim agus na bearta cánach agus custam...
Response (gle)	Fógra tábhachtach: An t-eolas is déanaí faoi sheirbhísí na gCoimisinéirí Ioncaim agus bearta cánach agus custam...

Table 7: BLEU gle2eng Example

NQ-eng (Kwiatkowski et al., 2019) evaluates world knowledge via open-ended generation (5-shot). Success is measured by an exact match. Table 8 demonstrates a failure case. Although the model correctly identifies the date ("October 24"), it fails to specify the year required by the reference.

Question	When did Taylor Swift's first album release?
Response	October 24
Reference	["October 24, 2006", "2005"]
Exact Match	**No**

Table 8: NQ-eng Example (Exact Match failure).

4.3.2 Benchmark Results

The success on both English and Irish benchmarks after CPT demonstrates the effectiveness of bilingual pre-training. Qomhrá outperforms the other two 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.

Qomhrá outperformed UCCIX by 29.2% in Irish as opposed to 44.4% in English on the IQA benchmark. The information from these questions is the same so language capability must be the factor that caused Qomhrá to outperform UCCIX, indicating improved English performance retention for Qomhrá, likely due to its higher proportion of English data at 25% compared to UCCIX's 1% . Qomhrá did not improve performance with the second epoch of training. This is in line with expectations that base models saturate after one epoch of CPT (Lu et al., 2025). Therefore in future development of the model, CPT would only be trained for one epoch to prevent redundant compute expenditure.

4.4 Human Feedback

Contrary to the instruction tuning inter-annotator agreement, the LLM and the human showed near-perfect annotation alignment. We hypothesise that this reflects the level of contrast between candidates in the annotations, i.e., LLMs can be used to substitute human annotators in low-resource languages for less complex tasks.

4.5 Error Analysis

We examine each benchmark individually to analyze task-specific errors and compare Qomhrá-2e-CPT against Qomhrá-Instruct to measure the impact of instruction tuning.

4.5.1 Cloze-gle

We represent the task of assigning the correct pronoun as a classification problem and can thus visualise errors with a confusion matrix (CM) displayed in Figure 2. We extract the predicted pronoun from Qomhrá's answer and the actual pronoun from the target sentence.

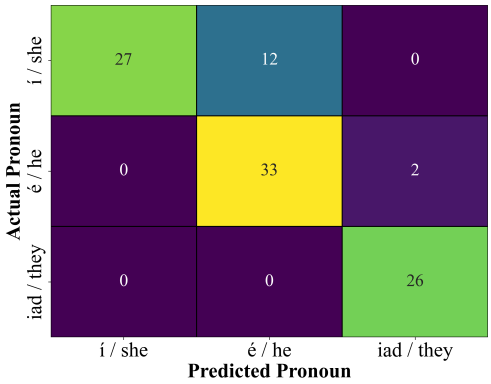

Figure 2: Qomhrá-2e-CPT, Cloze-gle Confusion Matrix: comparing the expected vs predicted pronoun. No significant difference with Qomhrá-Instruct CM.

In 100% of the errors, we note that the evaluated pronoun precedes the noun that it references, suggesting that word order plays a role. This can be attributed to Qomhrá decoding the sentence from left to right, masking the relevant gender information needed to select the correct pronoun making the task ambiguous.

In these cases of ambiguity, Qomhrá overwhelmingly selects the masculine pronoun *é* reflecting bias in the training data as the sentence probability is not penalized by the incorrect agreement. This bias is documented in neural machine translation (NMT) for similar cases where the gender of the author is not known (Vanmassenhove et al., 2018).

Qomhrá-2e-CPT and Qomhrá-Instruct agree for 92% of the Cloze-gle questions. Of the questions where they differ, no gender information proceeds

the pronoun. Therefore, Qomhrá-Instruct maintains the same bias towards the male pronoun. This is unsurprising as instruction tuning is not intended to correct existing linguistic biases.

4.5.2 SIB-gle

For the SIB topic classification benchmark, Qomhrá demonstrated high accuracy so we analyse the misclassified categories where the rate of error was over 10%. Qomhrá-2e-CPT misclassified Entertainment as Science/Technology in 31.25% of cases. This can be explained by words associated with technology being present in entertainment, such as video games, television, and film cameras. Secondly, Travel was misclassified as Geography in 12.82% instances where shared concepts such as weather, transport, landmarks and regional laws lead to confusion. Geography has no misclassification in Qomhrá-2e-CPT but 12.77% in Qomhrá-Instruct. These errors are examples of a lexical overlap heuristic (McCoy et al., 2019), where Qomhrá depends on vocabulary instead of more complex language understanding. This is further exacerbated by instruction tuning.

4.5.3 IQA

The IQA-eng and IQA-gle dataset test Irish cultural knowledge in both Irish and English. CPT is effective at significantly improving Qomhrá's performance in Irish, however, it still performs consistently stronger when tested in English. This reflects the stronger capabilities of knowledge retrieval in English. This is unsurprising as the Qwen3 pretraining corpus, where the model obtains the vast majority of its factual knowledge is in English and not Irish. Therefore, future work should aim to not only instil language capabilities, but also cultural knowledge.

4.5.4 BLEU

Upon manual inspection of Qomhrá's underperformance on machine translation (BLEU), we observe that Qomhrá-2e-CPT (not instruction tuned) fails to output the stop token. When evaluated, Qomhrá first sees few-shot examples of translations where the sentence is tagged by language as either Irish (*Gaeilge*) or English (*Béarla*). After translating the sentence being tested, Qomhrá mimics this tagging and continues to generate other example translations. We can therefore uncover Qomhrá's true translation performance by using ["Gaeilge:", "Irish:", "Béarla:", "English:"] as stop tokens.

Model	Direction	Raw	Cleaned	↑ Δ
CPT	gle2eng	5.19	22.25	17.06
	eng2gle	3.63	21.75	18.12
Instruct	gle2eng	11.66	11.66	0.00
	eng2gle	7.70	7.71	0.01

Table 9: BLEU performance of Qomhrá-2e-CPT and Qomhrá-Instruct before (Raw) and after (Cleaned) stop-token extraction. Where ↑ Δ is the positive change in BLEU score.

Firstly, we note that Qomhrá's true translation capability is much stronger than previously measured. The previous trend seen in Table 3 of scores improving following instruction tuning proves to be misleading. In fact, while instruction tuning mitigates the failure to stop as displayed in the low ↑ Δ, translation capabilities drop significantly. Before instruction tuning, Qomhrá was the only model that demonstrated equal performance when translating to Irish or English. However, instruction tuning introduces bias toward English generation displayed in Table 9. This indicates that instruction tuning on machine translated data successfully teaches the model to answer questions but significantly degrades the model's low-resource language capabilities.

4.5.5 NQ

Similar to the BLEU benchmark, manual inspection of the NQ trivia benchmark showed that Qomhrá-2e-CPT fails to output the stop token. Equally, the benchmarks in 3 showed that instruction tuning improved performance. However, table 10 shows that relaxing the constraint from exact match to the target answer being present in Qomhrá's response tells a different story.

Model	Strict	Relaxed	↑ Δ
CPT	3.55	28.81	25.26
Instruct	12.69	23.38	10.69

Table 10: Natural Questions (NQ) performance comparison between CPT and Instruct when the **strict** exact match condition is **relaxed**.

We see that once again, while the model improves on evaluation, it actually degrades on underlying performance. This analysis highlights the importance of not stopping evaluation once a benchmark metric is obtained but analysing the errors and

evaluation dynamics to provide a deeper and more transparent evaluation of the model.

5 Conclusion

Our paper presents Qomhrá, a bilingual Irish and English LLM, outlining a full CPT and instruction-tuning pipeline with human preference data synthesis. Our increased English language proportion in the CPT training highlights its necessity when training systems adapted to real-world context, where the low-resource and high-resource language coexist.

We provide the first human evaluation comparison of existing LLMs for Irish. This guides anyone seeking to integrate LLMs into language technology for Irish. In creating a 30K parallel instruction dataset and a 1K human preference dataset, we contribute a validated labelled data at a scale that can assist others in developing Irish-language LLMs.

Furthermore, our inter-annotator analysis showed that while L2 speakers and LLMs failed to align with an L1 speaker for complex annotation, an LLM can create a high-quality DPO dataset ($\kappa = 0.978$ alignment), if candidate contrast is high. This provides a useful heuristic for others considering data synthesis using existing LLMs to distill knowledge.

Our results establish key elements useful for developing chatbots for Irish and we hope that work of this nature can spur and aid others in developing LLMs for Irish and other low-resource languages.

Limitations

Though the closed-weight LLMs evaluated at the time of research were the state of the art, model providers have released new iterations since then, motivating updated comparisons. Due to compute limitations, our work is limited to the 8B model scale and a focus on Irish and English. Future work should include more languages and base models to support the generalization power of results.

Equally, only one open-weight model: Qwen3 served as the base model, where in future, we recommend independent empirical testing of available open-weight models specific to domain adaption, i.e., evaluating a proxy for Irish language performance before commencing the costly CPT, fine-tuning pipeline.

Regarding our annotator alignment findings, they are limited by access to the inclusion of only one L1 speaker, which encourages future work to collaborate at a larger scale with the low-resource language community.

Furthermore, evaluation does not address cultural alignment, code-switching, or dialectal variation, which are key functionalities that should be addressed in future work.

Ethical Considerations

Developing human language technology imposes ethical considerations concerning the human interfacing with the technology. As Qomhrá is primarily trained on web-crawled data, content can potentially be harmful. Future research should involve more rigorous pre-processing to remove this content. Qomhrá is primarily released for research purposes and users of Qomhrá should be mindful of potential harms in its generated outputs.

LLM CPT requires significant compute power which contributes to greenhouse gas emissions, as such it is important to quantify our impact. The power usage of an Nvidia H100 GPU is 700W. We trained for 44.196 hours on 2 GPUs. The carbon intensity was approximately 200 gCO_2/kWh (EirGrid, 2025), which gives a total of 12.37kg CO_2. This can be understood relative to the average of 10.4 tonnes of CO_2 per capita in Ireland (Central Statistics Office of Ireland, 2025). Efficient corpus sampling methods can be used in future work for more efficient training.

6 Acknowledgements

This research was carried out as part of the ABAIR project, supported by the Dept. of Rural and Community Development and the Gaeltacht, with funding from the National Lottery, as part of the Straitéis 20 Bliain don Ghaeilge.

We would like to thank CloudCIX Limited for the support of computing resources on their NVIDIA HGX/H100 cluster.

This publication has emanated from research supported in part by Research Ireland under Grant [18/CRT/6223].

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

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.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

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.

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.

Central Statistics Office of Ireland. 2025. Environmental indicators ireland 2025 - global context and climate.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2475–2485.

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.

EirGrid. 2025. CO_2 Intensity, Ireland.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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.

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

Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2022. Automatic speech recognition for Irish: the ABAIR-ÉIST system. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 47–51, Marseille, France. European Language Resources Association.

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.

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.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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.

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.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

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.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3003–3008.

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*.

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.