GaelEval: Benchmarking LLM Performance for Scottish Gaelic

Peter Devine,[1] **William Lamb,**[1] **Beatrice Alex,**[1] **Ignatius Ezeani,**[2] **Dawn Knight,**[3]
Mícheál J. Ó Meachair,[4] **Paul Rayson,**[2] **Martin Wynne**[5]
[1]University of Edinburgh, [2]Lancaster University, [3]University of Cardiff,
[4]Dublin City University, [5]University of Oxford
{pdevine2, w.lamb, b.alex}@ed.ac.uk, {i.ezeani, p.rayson}@lancaster.ac.uk,
knightd5@cardiff.ac.uk, micheal.omeachair@dcu.ie, martin.wynne@ling-phil.ox.ac.uk

Abstract

Multilingual large language models (LLMs) often exhibit emergent 'shadow' capabilities in languages without official support, yet their performance on these languages remains uneven and under-measured. This is particularly acute for morphosyntactically rich minority languages such as Scottish Gaelic, where translation benchmarks fail to capture structural competence. We introduce **GaelEval**, the first multi-dimensional benchmark for Gaelic, comprising: (i) an expert-authored morphosyntactic MCQA task; (ii) a culturally grounded translation benchmark and (iii) a large-scale cultural knowledge Q&A task. Evaluating 19 LLMs against a fluent-speaker human baseline ($n = 30$), we find that Gemini 3 Pro Preview achieves 83.3% accuracy on the linguistic task, surpassing the human baseline (78.1%). Proprietary models consistently outperform open-weight systems, and in-language (Gaelic) prompting yields a small but stable advantage ($+2.4\%$). On the cultural task, leading models exceed 90% accuracy, though most systems perform worse under Gaelic prompting and absolute scores are inflated relative to the manual benchmark. Overall, GaelEval reveals that frontier models achieve above-human performance on several dimensions of Gaelic grammar, demonstrates the effect of Gaelic prompting and shows a consistent performance gap favouring proprietary over open-weight models.

Keywords: benchmarking, multilingual evaluation, large language models, morphologically rich languages, Scottish Gaelic

1. Introduction

Although most large language models (LLMs) officially support a small fraction of the approximately 7,000 human languages spoken worldwide, they display emergent 'shadow' capacities in many more. For instance, OpenAI advertises support for 59 languages in ChatGPT,[1] none of which belong to the Celtic family (e.g. Irish, Welsh and Scottish Gaelic). Despite this, the system processes and generates text in every Celtic language. The distinction between official and *de facto* support raises a methodological challenge: as coverage expands and model varieties diversify, establishing robust evaluation frameworks becomes crucial for both official languages and the minority languages they nevertheless represent.

The current evaluation landscape is markedly skewed. High-resource languages like English benefit from a self-reinforcing ecosystem of training corpora and mature benchmarks. In contrast, low-resource languages suffer from sparse training data (Joshi et al., 2020) and little or no evaluation resources (Romanou et al., 2024). Even where benchmarks exist, they rarely include human baselines (Assadi et al., 2025), making it impossible to ascertain whether a model's output follows a given community's sociolinguistic norms or not. Furthermore, the English-dominance of these models risks processing the world's cultural-linguistic mosaic through an Anglocentric lens. Without objective measurement, academics and language communities alike cannot determine if an LLM should be explored or eschewed.

Scottish Gaelic ('Gaelic') epitomises these challenges. Ranking 104th in Common Crawl accessibility,[2] Gaelic occupies the digital margins, yet it possesses a rich morphosyntax that defies the only benchmarks available for it: the surface-level translation based FLORES-200 (Goyal et al., 2022) and BritEval (BritLLM, 2026). These benchmarks do not capture whether a model is truly 'Gaelic-conversant' or merely performing a high-dimensional translation of English concepts.

To address this gap, we present **GaelEval**: a targeted, multi-dimensional evaluation suite that moves beyond surface equivalence toward deeper morphosyntactic and culturally grounded competence. Our framework includes three distinct tasks:

1. **Linguistic Competence**: A multiple-choice question answering (MCQA) task comprising 120-questions and probing fine-grained grammatical and idiomatic usage.

2. **Translation**: A rigorous assessment using BLEU and chrF metrics against hand-translated gold labels.

[1] https://help.openai.com/en/articles/8357869-how-to-change-your-language-setting-in-chatgpt. Accessed 22 Feb 2026.

[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages. Accessed 21 Feb 2026.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 73–85
11 May 2026.

3. **Cultural Understanding**: A culturally grounded Q&A task (1,087 questions) derived from pedagogical content produced by fluent speakers.

We evaluate 19 contemporary LLMs (14 proprietary; 5 open-weight), providing the first systematic comparison of LLM performance for Scottish Gaelic. Gemini 3 Pro Preview leads overall and surpasses the fluent-speaker baseline on the linguistic competence task.

Our principal contributions are:

- **GaelEval**, the first multi-dimensional benchmark for Scottish Gaelic, spanning morphosyntax, translation and culturally grounded knowledge;
- the first human baseline for Gaelic LLM evaluation ($n = 30$);
- evidence that Gemini 3 Pro Preview exceeds the fluent-speaker mean on a controlled morphosyntactic task;
- a consistent aggregate advantage for in-language (Gaelic) prompting for the morphosyntactic task; and
- quantification of the performance gap between frontier proprietary and open-weight models in a minority-language setting.

In what follows, we review related work (§2), describe our design and evaluation methodology (§3), present empirical results (§4) and conclude, with proposed directions for future work (§5).

2. Related Work

Multilingual LLM Evaluation Frameworks Large-scale multilingual benchmarks are central to evaluating LLM capabilities across languages. MMLU (Hendrycks et al., 2020) introduced a widely used multiple-choice framework for knowledge-intensive reasoning in English. Global MMLU (Singh et al., 2025) extended this paradigm cross-lingually, largely via translation of English-source materials. FLORES-200 (Goyal et al., 2022; NLLB Team et al., 2024) expanded coverage to 200+ languages, including Gaelic, but evaluates only machine translation (MT). BritEval (BritLLM, 2026) consists of 3 major English benchmarks translated into 4 languages from Britain and Ireland, including Gaelic. XTREME-UP (Ruder et al., 2023) incorporates additional low-resource tasks (e.g., transliteration, OCR), while INCLUDE (Romanou et al., 2024) departs from translation-based design by constructing question answering benchmarks from native regional exam materials.

Many multilingual benchmarks rely heavily on translation or adaptation from English-centric datasets. While valuable, this approach underrepresents language-specific morphosyntax, culturally grounded knowledge, and idiomatic usages that resist direct translation (e.g. that the colour of grass in Gaelic is *gorm* 'lit. blue', not *uaine* 'lit. green'). For morphologically rich languages such as Gaelic, translation-based evaluation also is unlikely to capture fine-grained inflectional contrasts or edge cases that distinguish structural competence from superficial word recognition. To our knowledge, beyond BritEval, FLORES-200 and related benchmarks (e.g., SIB-200; Adelani et al., 2024), no large-scale Gaelic evaluation suite exists.

Low-Resource and Morphologically Rich Language Evaluation Recent work addresses the challenges of evaluating LLMs on low-resource and morphologically complex languages, including tokenisation and pattern extrapolation (Xia et al., 2025). IndicGenBench (Singh et al., 2024) covers 29 Indic languages using human-curated parallel data; AfriQA (Ogundepo et al., 2023) introduces question answering for African languages; and TurkBench (Toraman et al., 2026) evaluates Turkish across 21 subtasks. Xia et al. (2025) further propose a cross-lingual benchmark spanning Cantonese, Japanese and Turkish, combining human evaluation with automated metrics across diverse tasks. Irish-BLiMP evaluates LLMs on Irish linguistic knowledge using 1020 minimal pairs and provides a human baseline (McGiff et al., 2025). Collectively, these efforts highlight the need to evaluate LLMs on morphologically rich, low-resource languages. We extend this line of work by directly assessing model competence in Gaelic morphosyntax and non-compositional usage.

Culturally and Linguistically Informed Evaluation A growing literature argues that linguistic competence cannot be evaluated independently from cultural knowledge. Tao et al. (2024) document systematic bias toward English-speaking contexts in ostensibly multilingual LLMs. Relatedly, multilingual models have been shown to process non-English inputs through English-dominant representational pathways (Papadimitriou et al., 2023; Wendler et al., 2024), raising concerns about whether these systems encode language-specific structures or just rely on Anglocentric priors.

Recent benchmarks increasingly integrate cultural and linguistic evaluation. For example, ProverbEval (Azime et al., 2025) assesses Ethiopian languages (and English) through proverb interpretation, requiring both morphosyntactic competence and culturally grounded reasoning. Knowledge-

grounded benchmarks similarly test community-specific factual knowledge in domains such as food, holidays and social practices (Myung et al., 2025). Importantly, Myung et al. (2025) show that in-language prompting benefits medium- and high-resource languages, whereas low-resource languages often perform better under English prompting. This resource-sensitive pattern motivates our evaluation under both English and Gaelic prompt conditions to test whether similar asymmetries arise for Gaelic.

LLM-Assisted Benchmark Construction LLM-assisted benchmark generation offers a practical solution when human-curated datasets are scarce (Perez et al., 2023; Anwar et al., 2026). Prior work has used LLMs to extract cultural knowledge from large corpora such as C4 (Nguyen et al., 2023) and TikTok (Shi et al., 2024), derive culturally grounded Q&A from web scrapes (Wang et al., 2024) and Wikipedia (Fung et al., 2024), and generate multilingual evaluation data across 13 languages (Zhao et al., 2025). Although LLMs typically perform worse on low-resource languages, raising concerns about synthetic benchmark quality, manual analysis of 10,000 generated instructions in 13 Indic languages found over 99.7% to be of high or moderate quality (Chitale et al., 2025).

In our setting, automated generation was required for scale. To reduce associated risks, we applied structured filtering and answerability scoring (§3.1.3), discarding items below predefined thresholds. While not a substitute for native-speaker validation, this procedure provides a systematic safeguard against noise and incoherence.

3. Benchmark Design and Evaluation Methods

In this section, we describe the design of **GaelEval** and our evaluation methods. Unlike translation-based frameworks such as FLORES-200, GaelEval integrates an expert-designed morphosyntactic MCQA task with culturally grounded Gaelic-source texts for translation and Q&A evaluation.

3.1. Tasks

3.1.1. Linguistic Competence

We define *linguistic competence* as the ability to select grammatically and idiomatically appropriate forms in controlled morphosyntactic contexts. The 120-item MCQ set was designed by a Gaelic domain expert, who used a recent grammar (Lamb, 2024) to identify grammatical edge cases (e.g. long-distance relativisation) and constructions resistant to literal translation from English.[3] Items span 16 grammatical categories (Table 1), with nominal morphology the largest (17 items; 14.6%). The design prioritised breadth across grammatical domains (e.g. case marking, agreement, idiomatic conventions) over depth within individual micro-phenomena (e.g. the feminine singular genitive), while keeping the task manageable for human participants. The MCQs were not publicly available prior to evaluation, ensuring zero-shot conditions.

Category	N
Nominal morphology	17
Adjectives	11
Verbal noun cores	12
Formulaic expressions	10
Questions and tags	10
Prepositions	9
Pronouns and anaphor resolution	9
Tense Aspect Modality (TAM) system	7
Impersonals and passives	7
Adverbials	5
Conjunctions and particles	5
Relative clauses	5
Clefts and focussing expressions	4
Colours	3
Determiners	3
Numerals	3
Total	**120**

Table 1: Distribution of MCQs across principal grammatical categories.

Each question contained a single gap and was designed to have one unambiguous correct answer. For example, the following question asks for the feminine singular basic definite form of the noun *fuil* 'blood', where b. is the correct answer.

```
Chunnaic Màiri ___.
('Mary saw ___.')
a. am fuil
b. an fhuil
c. am fhuil
d. na fala
```

Distractors were constructed to be linguistically plausible and, as much as possible, attested in contemporary usage. This ensured that a successful choice required grammatical discrimination rather than mere lexical recognition. The placement of the correct answer was varied during preparation, but the task was issued in a fixed order across all iterations to ensure consistency for humans and models.

Human participants were recruited via convenience sampling on social media (Facebook and

[3] Our approach is inspired by an unpublished Irish-language MCQA study by Joseph McInerney.

	FLUENT		NON-FLUENT		
Onset	**Near-/Native**	**Adv**	**Upp-Int**	**Int**	**Sum**
0–4	12	0	0	0	12
12–18	2	2	1	0	5
18–30	7	5	3	1	16
30+	2	0	0	0	2
Total	**23**	**7**	**4**	**1**	**35**

Table 2: Participant distribution by age of first exposure to Gaelic (Onset), self-reported proficiency level ($N = 35$) and fluency grouping. 'Fluent' combines near-/native and advanced.

LinkedIn), yielding a voluntary, non-representative sample of Gaelic speakers ($N = 35$; $n = 30$ after filtering). The task was administered online via Qualtrics. Following informed consent and instructions in English, for accessibility, responses were collected anonymously along with self-reported proficiency and age of acquisition ("Onset"). Proficiency levels were: Near-/Native, Advanced (fluent in most contexts), Upper intermediate (comfortable in most discussions), and Intermediate (everyday conversational ability).

For benchmarking, we combined the Near-/Native and Advanced groups as 'Fluent' to approximate stable adult competence. While native speakers typically acquire Gaelic in childhood, advanced and near-native speakers often receive formal instruction. This may increase familiarity with the prescriptive grammatical forms targeted in the MCQA.

The distribution of participants by onset and proficiency is shown in Table 2. Approximately one third (12/35) identified as Near-/Native and reported acquiring Gaelic before age five, consistent with socialisation in a Gaelic-speaking home. We refer to this group as 'native' in Table 5.

Mean task duration was 45.7 minutes (median = 33.75; range = 12.5–165.2). Questions were presented in fixed order, and participants were required to select a single response per item. Missing responses were scored as incorrect; submissions with more than 10 missing items were excluded.

Following task administration, a few questions were flagged as potentially dialect-sensitive or admitting multiple acceptable responses. After review in light of documented Gaelic morphological variation (Adger, 2010; Iosad and Lamb, 2020), three items (IDs 12, 21, 48) were excluded. All reported results, therefore, are based on 120 questions versus the original 123.

We input the MCQs to the models listed in Table 3 (see §3.2) with each item evaluated in a single-turn call. For each call, the prompt comprised a fixed system instruction and a user message containing the question sentence with a single blank and the full list of answer options (see §6.1). Models were instructed to return only the text of the correct option (e.g. `b. an fhuil`), without explanation or additional punctuation, and short examples were included to enforce this format.

Decoding parameters were left at model defaults. Responses were scored by exact string match after trimming whitespace; outputs beginning with `Error:` were logged as API failures. To mitigate transient failures and rate limits, calls were retried with exponential back-off (up to five attempts), and per-item outputs and correctness flags were stored in JSONL format.

3.1.2. Culturally Relevant Translation

To construct the translation task, we collected parallel English and Gaelic transcripts of the Gaelic learning podcast *An Litir Bheag* ('The Little Letter') from LearnGaelic.scot. As these transcripts are professionally translated, we treat them as gold references for English–Gaelic evaluation. Produced by a fluent Gaelic speaker for intermediate and advanced learners, the episodes frequently address culturally salient topics, making them a relevant resource for assessing LLM performance on MT.

We downloaded all available episodes of *An Litir Bheag* with both English and Gaelic transcripts at the time of writing (episodes 154–1076). Five episodes were excluded due to failed mp3 downloads. (We had initially intended to compile a parallel corpus including Gaelic audio to support ASR evaluation.) The final dataset comprises 918 English–Gaelic parallel transcripts.

On manual inspection, we found errors in how the podcast producers had published some of the podcasts, and so we performed filtering to ensure general data quality. First, we found that some transcripts had their identifying language reversed (i.e. Gaelic vs English, and vice-versa), so we applied a text language identification model, OpenLID v2 (Burchell et al., 2023), to all transcripts and manually removed those that showed switched languages.

We also identified cases where transcripts did not correspond to the correct podcast or failed to align as parallel pairs. To detect such mismatches systematically, we translated all Gaelic transcripts into English using GPT-5.2 and computed BLEU scores against the paired English versions using SacreBLEU (Post, 2018). Episodes were sorted by lowest BLEU score to identify likely mismatches. Following automated filtering and manual inspection, 10 episodes were removed, yielding a final dataset of 908 parallel transcripts.

Finally, we downloaded episode subject metadata from the LearnGaelic.scot website and aligned it with the episode transcripts. This enabled us to

Open-weight: DeepSeek R1, GLM 4.7, GPT OSS 120B, GPT OSS 20B, Llama 4 Maverick. **Closed-weight:** Claude Haiku 4.5, Claude Opus 4.6, Gemini 2.5 Flash, Gemini 3 Flash Preview, Gemini 3 Pro Preview, GPT-4.1, GPT-4.1 Mini, GPT-4.1 Nano, GPT-4o, GPT-4o Mini, GPT-5, GPT-5 Mini, GPT-5 Nano, GPT-5.2

Table 3: Models evaluated in GaelEval.

classify each episode according to one of the following thematic categories: Folklore, Gaelic language, History, Nature, Pastimes, People or Places.

3.1.3. Q&A Task on Cultural Understanding

Alongside our parallel transcripts, we also generated a MCQ set to assess the LLMs' level of Gaelic understanding and knowledge. We first instructed GPT-5.2 to rate the cultural significance of each episode's transcripts from 1-5 (system messages are detailed in the Appendix 6.2). For this task, we removed any episode rated less than 4 to filter transcripts unrelated to general knowledge of Gaelic culture (e.g. autobiographic episodes). Following this procedure, we maintained 713 episodes out of the original 908.

We then instructed GPT-5.2 to generate between 1 and 10 general knowledge-style Q&A pairs per episode based on the episode transcripts in Gaelic, alongside English translations of each question and answer. This yielded 6,802 Q&A pairs.

Occasionally, generated questions referred to transcript-specific details despite instructions to avoid contextual dependence, rendering them unsuitable as stand-alone general knowledge items. We therefore used GPT-5.2 to assign an 'answerability' score (1–5) to each Gaelic question and its English translation, where 5 denotes a fully self-contained general-knowledge item and 1 indicates contextual dependence. Questions scoring below 4 in either language were excluded. This filtering yielded a final set of 1,087 questions drawn from 440 unique episodes, a subset of which was manually verified by a Gaelic domain expert.

Finally, we re-input the transcripts to GPT-5.2 together with the generated questions and answers, instructing it to produce three plausible distractors per item. English translations were also generated for each distractor. This yielded 1,087 multiple-choice questions, each comprising one correct answer and three distractors. Answer options were randomly shuffled and labelled A–D, to prevent positional bias and ensure consistent single-letter responses from models.

3.2. LLM Models Evaluated

We evaluated 19 models (Table 3) across three tasks: linguistic competence, translation and cultural understanding, spanning both open- and closed-weight systems. Models were accessed via the OpenAI, Anthropic, Google AI Studio and Together AI endpoints, with batch processing used to reduce costs. Responses were constrained to a predefined JSON schema (single key–value pair) to minimise explanatory output preceding the answer.

3.3. Evaluation Metrics

For the MCQA tasks, we report accuracy, defined as the percentage of outputs that both conformed to the required JSON schema (see Section 3.2) and contained the correct answer. For translation, we report BLEU (Papineni et al., 2002) and chrF (Popović, 2015).

4. Results

4.1. Linguistic Competence Task

As detailed in §3.1.1, we deployed a 120-item MCQ set to assess models' Gaelic linguistic competence. We also administered the same task to human participants ($N = 35$). Table 4 reports model accuracy under two prompt conditions (Gaelic and English system messages: see §6.1), alongside the performance of fluent speakers ($n = 30$); we exclude intermediate learners to approximate stable adult competence.

Gemini 3 Pro Preview achieves the strongest overall results, scoring 83.3% under Gaelic prompting and 80.0% under English prompting. Its Gaelic-prompted score is significantly higher than the fluent-speaker mean of 78.1% (95% CI: 73.9%-82.2%; $p < 0.05$). This contrasts with comparable work on Irish, where fluent speakers ($n = 3$) outperformed all evaluated LLMs (McGiff et al., 2025). Among OpenAI's models, GPT-5 performs best (69.2% Gaelic; 71.7% English), while Claude Opus 4.6 leads the Anthropic systems (59.2% Gaelic; 50.8% English), though it remains more than 24 percentage points below Gemini 3 Pro Preview under Gaelic prompting.

Among open-weight systems, DeepSeek R1 (45.0% Gaelic; 42.5% English) and Llama 4 Maverick (45.0% Gaelic; 40.0% English) performed best, yet remained approximately 38 percentage points below Gemini 3 Pro Preview. Both models were only modestly above the 25% chance baseline, consistent with results reported for other morphologically rich low-resource languages (Etxaniz et al., 2024; McGiff et al., 2025). This gap may reflect differences in training data scale and

composition between open- and closed-weight systems. A plausible factor in Google's favour is its long-standing support for Scottish Gaelic in Google Translate (since 2016) (BBC, 2016), and the associated accumulation of Gaelic data.

We observe a modest aggregate advantage for Gaelic prompting. Averaged across models (excluding humans), Gaelic system messages outperform English ones by 2.4 percentage points (46.8% vs. 44.4%). Although the effect varies by model family, the overall pattern indicates a small but consistent in-language advantage. This pattern is interesting in light of Myung et al. (2025), who find that in-language prompting benefits medium- and high-resource languages but not typically low-resource ones. At the same time, the advantage is small and contrasts to our findings for the cultural understanding task (see §4.3, Table 7).

Notwithstanding the small sample size (see §3.1.1), advanced speakers – most of whom reported acquiring Gaelic between ages 18 and 30 – outperformed native speakers on this task (82.6% vs 70.9%; Table 5). Feasibly, some tested phenomena involve formal registers or infrequent edge cases more likely to be acquired through structured learning than everyday usage. Within our sample, self-reported fluency therefore does not fully align with prescriptive grammatical knowledge.

Within model families, we observe patterns consistent with known scaling-law behaviour. Performance improvements have been shown to follow predictable power-law trends as training compute, dataset size and parameter count increase proportionately (Kaplan et al., 2020; Hoffmann et al., 2022). Table 4 reflects this tendency: more recent flagship models generally outperform earlier versions within the same family (e.g. Gemini 3 Pro: 83.3% vs Gemini 2.5 Flash: 61.7% in the Gaelic condition). Similar intra-family gains are observed among OpenAI and Anthropic models.

Smaller 'Mini' and 'Nano' variants – designed to reduce parameter count and inference cost through compression and distillation – consistently underperform their corresponding flagship models (e.g. GPT-5: 69.2% vs GPT-5 Mini: 47.5% under Gaelic prompting), reflecting the expected trade-off between efficiency and representational capacity. While recent work shows that smaller models can achieve strong performance in low-resource contexts when specifically adapted (Etxaniz et al., 2024), our results indicate that, without such adaptation, reduced-capacity variants perform well below large proprietary models for Gaelic.

Overall, under Gaelic prompting, Gemini 3 Pro Preview demonstrates linguistic competence in Gaelic exceeding that of fluent speakers, with GPT-5's English-prompted performance slightly under the human baseline. Although the extent to which this task generalises to broader real-world text production remains unclear, we observe comparable performance patterns in the translation and cultural understanding tasks discussed below, supporting its construct validity.

Model / Group	Gaelic	English	Δ G-E
Gemini 3 Pro Prev	**83.3**	**80.0**	3.3
Gemini 3 Flash Prev	79.2	77.5	1.7
Humans (Fluent)	n/a	78.1	–
GPT-5	69.2	71.7	-2.5
Gemini 2.5 Flash	61.7	62.5	-0.8
Claude Opus 4.6	59.2	50.8	8.3
GPT-4o	50.0	46.7	3.3
Claude Haiku 4.5	47.5	43.3	4.2
GPT-5 Mini	47.5	41.7	5.8
DeepSeek R1	45.0	42.5	2.5
Llama 4 Maverick	45.0	40.0	5.0
GPT-5.2	43.3	40.0	3.3
GPT-4.1	42.5	42.5	0.0
GPT-5 Nano	36.7	30.8	5.8
GPT-4o Mini	34.2	30.0	4.2
GPT OSS 120B	33.3	27.5	5.8
GLM 4.7	32.5	26.7	5.8
GPT-4.1 Nano	30.0	31.7	-1.7
GPT OSS 20B	29.2	22.5	6.7
GPT-4.1 Mini	24.2	27.5	-3.3
Mean (models)	46.8	44.4	2.4

Table 4: MCQA accuracy (%) under Gaelic and English prompting. Δ indicates Gaelic minus English performance. Human baseline shown for fluent speakers (n=30).

Finally, Figure 1 disaggregates accuracy by grammatical category (cf. Table 1), comparing the fluent-speaker baseline with the nine highest-performing models. Two contrasts are evident. First, models underperform humans on interactional and idiomatic material – most clearly in questions/tags (Human = 0.85; Gemini 3 Pro = 0.50; GPT-5 = 0.30) and formulaic expressions (Human = 0.82; top models ≈ 0.40–0.60). Second, several models match or exceed the human mean in more structural domains, including determiners, pronouns and anaphor resolution, and relative clauses (e.g. REL: Human = 0.53 vs. Gemini 3 Pro ≈ 0.80). Given uneven category sizes, these differences are indicative rather than definitive. Nevertheless, if replicated with a larger, balanced MCQA, the pattern would suggest that current models handle codified morphosyntax more reliably than interactional discourse, with implications for conversational AI and CALL targeting everyday Gaelic.

4.2. Culturally Relevant Translation

Translation metrics are reported in Table 6. We observe consistent score degradation when translating into Gaelic (en→gd) versus English (gd→en).

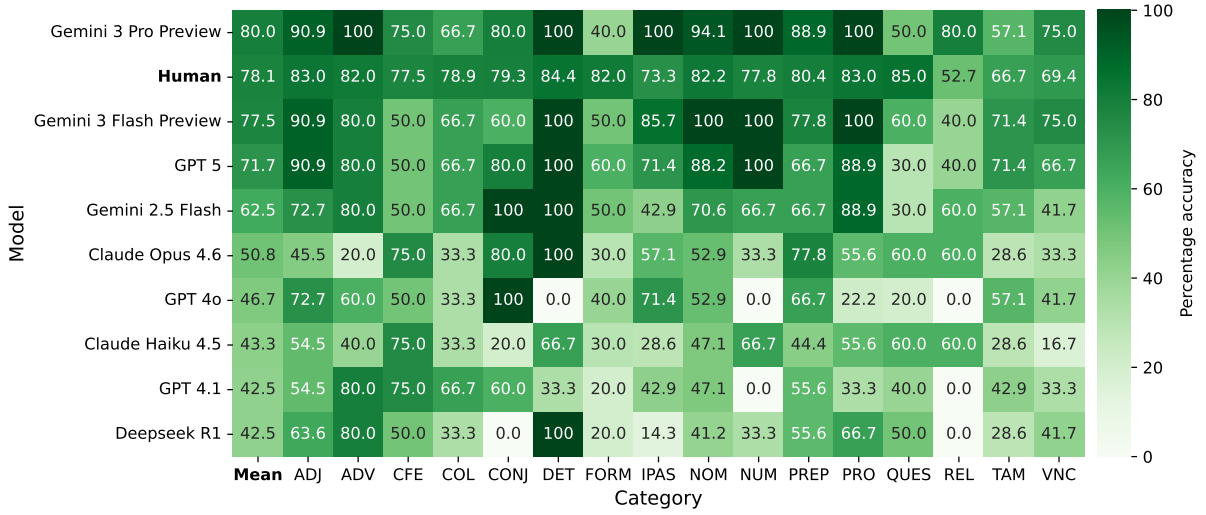

Figure 1: Linguistic competence accuracy by grammatical category for the human baseline and the top nine most-performant models (0–1 scale; darker = higher accuracy). ADJ = adjectives; ADV = adverbials; CFE = clefts and focussing expressions; COL = colours; CONJ = conjunctions and particles; DET = determiners; FORM = formulaic expressions; IPAS = impersonals and passives; NOM = nominal morphology; NUM = numerals; PREP = prepositions; PRO = pronouns and anaphor resolution; QUES = questions and tags; REL = relative clauses; TAM = Tense-Aspect-Modality system; VNC = verbal noun cores. Means are cross-category and so differ from those in Table 4. English prompting conditions used.

Level	N	Mean	Max	Min
Near-/Native	23	76.7	98.3	53.3
Native only	12	70.9	85.0	53.3
Advanced	7	**82.6**	97.5	63.3
Upper Intermediate	4	73.1	79.2	67.5
Intermediate	1	41.7	41.7	41.7
Fluent	30	78.1	98.3	53.3

Table 5: Human accuracy (%) on the 120-item MCQA by self-reported proficiency. 'Fluent' includes near-/native and advanced speakers.

This corroborates prior work showing that generation in a low-resource language imposes greater representational demands than comprehension mapped back to a high-resource language such as English (Goyal et al., 2022).

Gaelic to English MT largely tests how well a model projects low-resource inputs into English-dominant latent spaces and then generates in English. Conversely, English to Gaelic MT demands active production of morphologically complex and culturally-grounded forms. This asymmetry is especially pronounced in GPT 5 Mini, which drops over 20 BLEU points across directions. Gemini 3 Flash Preview is the exception, maintaining near symmetry (71.47 BLEU en→gd; 71.75 BLEU gd→en), suggesting a more balanced multilingual pre-training distribution.

Directional asymmetry is particularly pronounced among open-weight models. DeepSeek R1 attains the highest BLEU score for gd→en translation (75.28), marginally surpassing leading proprietary systems, but drops to 62.12 BLEU for en→gd generation. This pattern suggests that while mid-tier open-weight models can parse and comprehend Gaelic, they lack the generative capacity to produce morphologically fluent output. The effect is most extreme in GPT OSS 20B, which achieves 48.21 BLEU for gd→en yet collapses to 0.00 BLEU for en→gd, failing to generate valid Gaelic within the

	en → gd		gd → en	
	BLEU	chrF	BLEU	chrF
Gemini 3 Flash Preview	**71.47**	**79.07**	71.75	77.38
Gemini 3 Pro Preview	65.56	78.57	73.41	**78.20**
Gemini 2.5 Flash	65.53	74.72	74.36	77.58
GPT-4.1	65.41	74.62	66.60	72.77
Deepseek R1	62.12	71.84	**75.28**	77.73
GPT-5.2	61.19	72.16	71.02	76.18
GPT-5	56.94	70.27	70.27	75.82
Claude Opus 4.6	53.16	69.34	70.37	77.38
GPT-4o	52.39	67.83	68.73	74.80
GPT-5 Mini	49.19	60.93	69.58	73.34
Llama 4 Maverick	42.93	62.30	73.49	75.74
GPT-5 Nano	42.56	55.54	61.78	67.41
GPT-4.1 Mini	38.90	58.01	69.09	75.48
Claude Haiku 4.5	38.44	59.27	72.32	73.92
GPT OSS 120B	37.79	54.74	56.23	63.01
GPT-4o Mini	34.02	56.84	69.01	73.29
GPT-4.1 Nano	33.93	50.35	63.31	69.04
GPT OSS 20B	0.00	0.00	48.21	53.38
GLM 4.7	0.00	0.00	0.00	0.00

Table 6: BLEU and chrF scores for both English to Gaelic (en→gd) and Gaelic to English (gd→en) translation tasks.

Model	Gaelic	English	Δ G-E
Gemini 3 Flash Prev	**91.35**	**91.63**	-0.28
Gemini 3 Pro Prev	91.17	90.98	0.19
GPT-5	85.28	88.50	-3.22
Gemini 2.5 Flash	79.85	83.44	-3.59
Claude Opus 4.6	79.48	83.53	-4.05
GPT-5 Mini	71.48	80.04	-8.56
GPT-5.2	66.70	70.65	-3.95
GPT-4.1	66.24	74.70	-8.46
GPT-4o	63.85	73.41	-9.56
GLM 4.7	63.75	72.40	-8.65
Llama 4 Maverick	59.06	70.01	-10.95
DeepSeek R1	58.33	75.53	-17.20
GPT-5 Nano	54.37	70.29	-15.92
GPT OSS 120B	52.81	70.56	-17.75
GPT-4.1 Mini	47.93	65.59	-17.66
GPT-4o Mini	43.05	65.04	-21.99
GPT OSS 20B	42.41	57.04	-14.63
Claude Haiku 4.5	40.29	47.84	-7.55
GPT-4.1 Nano	34.87	51.61	-16.74
Mean	63.75	73.30	-9.55

Table 7: Accuracy (%) on the cultural knowledge Q&A task under Gaelic and English prompting. Δ indicates Gaelic minus English performance.

required JSON format.

We also note that GLM 4.7 was not able to form a single correctly formatted JSON response to any of our translation requests. This indicates that prompting some LLMs to process long passages of low-resource languages may degrade their ability to perform more basic tasks, such as JSON formatting.

Finally, divergence between word-level BLEU and character-level chrF highlight the challenges posed by Gaelic morphology. Gaelic's rich inflectional morphology means a model may retrieve the correct lemma yet miss the surface form that BLEU requires. Accordingly, chrF scores are consistently higher and less variable, particularly for en→gd translation, indicating difficulty with morphological realisation.

4.3. Q&A Task on Cultural Understanding

Table 7 reports accuracy on the Gaelic- and English-prompted versions of the cultural Q&A task. Gaelic-prompted performance ranges from 34.87% (GPT-4.1 Nano) to 91.35% (Gemini 3 Flash Preview), indicating substantial variation in cultural knowledge and reasoning. The strongest models – Gemini 3 Flash Preview, Gemini 3 Pro Preview, and GPT-5 – perform consistently well across both languages, with less than a 4-point gap between conditions.

In contrast to the linguistic competence task (see Table 4), most models perform worse under Gaelic prompting (mean Δ = -9.55), with the disparity widening among weaker systems. For example, GPT OSS 20B declines from 57.04% (English) to 42.41% (Gaelic), while GPT-4.1 Nano drops from 51.61% to 34.87%, approaching chance. This pattern aligns with prior findings that LLM representations for low-resource languages are weaker than for English (Goyal et al., 2022; Romanou et al., 2024; Singh et al., 2025).

Comparing these results with the hand-curated Linguistic Competence task reveals strong rank correlation across benchmarks. The top three models – Gemini 3 Flash Preview, Gemini 3 Pro Preview, and GPT-5 – retain the same ordering, and mid-tier open-weight systems (e.g., DeepSeek R1, Llama 4 Maverick) and smaller distilled variants exhibit similar relative positions.

However, absolute scores are inflated on the synthetically generated cultural QA task. The strongest models exceed 90% accuracy, whereas the linguistic competence MCQ peaks at 83.3%. This suggests that LLM-generated benchmarks may create easier evaluation conditions while nevertheless preserving relative performance rankings.

For low-resource languages lacking curated datasets, this result is encouraging: frontier models can generate synthetic cultural benchmarks from relevant authentic text, enabling scalable comparative evaluation. Although such datasets are not reliable as absolute measures of capability – scores may over-estimate fluency – they provide a low-cost and practical heuristic when creating a manual dataset would be prohibitive.

5. Conclusion

This paper introduces **GaelEval**, the first multi-dimensional benchmark for Scottish Gaelic, combining expert-authored morphosyntactic MCQA, culturally grounded translation and large-scale cultural knowledge evaluation. Across 19 contemporary LLMs, we observe variation in Gaelic competence, with proprietary systems outperforming open-weight models by a large margin. Notably, Gemini 3 Pro Preview exceeds the fluent-speaker baseline on the linguistic competence task, while translation and cultural Q&A results reveal directional asymmetries and consistent performance gaps between English- and Gaelic-prompted conditions.

Category-level analysis on the linguistic competence task suggests that current models handle codified morphosyntactic structure more reliably than interactional or formulaic usage. At the same time, rank correlation across the three tasks indicates that synthetic, LLM-generated cultural benchmarks can provide reliable evaluation signals, even if absolute scores may be inflated.

For minority languages lacking established eval-

uation infrastructure, GaelEval demonstrates that rigorous, language-specific benchmarking is both feasible and necessary. Future work will expand human validation, balance per-category item counts and incorporate further assessment of fluency and sociolinguistic appropriateness. Robust evaluation remains essential if 'shadow' LLM competence in low-resource languages is to be understood, trusted and responsibly deployed.

Data and Code Availability

To preserve the integrity and reusability of GaelEval as a zero-shot evaluation benchmark, the underlying dataset is not being released at this time. Public release likely would expose the data to web scraping bots, leading to its inclusion in future model pre-training corpora and compromising the benchmark's effectiveness. To support ongoing evaluation without exposing the data, a Gaelic LLM leaderboard is planned at https://eist.ac.uk. All code for prompt construction, API interaction, and evaluation will be released in the following public repository upon publication, ensuring transparency and reproducibility: https://github.com/Peter-Devine/gaeleval.

Ethical Considerations

Institutional ethical approval for this research was sought on 27 January 2026 and granted on 9 February 2026 by the Ethics Officer of the first author's host institution. The study involved minimal risk to participants and did not collect personally identifiable information beyond self-reported proficiency and age of acquisition.

Although no substantial personal or social risks are associated with this work, we acknowledge the environmental costs associated with large-scale model inference. While the evaluation tasks reported here required limited compute, the broader ecological impact of LLM development remains an important consideration.

Limitations

For the linguistic task, the human sample is small and non-representative (35 participants; 30 after filtering) and the MCQA is relatively short (c.f. McGiff et al., 2025) and unbalanced across categories. GPT-5.2 was used for data generation and filtering and was also evaluated, which may inflate its cultural Q&A performance due to self-preference bias (Wataoka et al., 2024; Xu et al., 2025). Moreover, the cultural Q&A benchmark is entirely LLM-generated, although a subset was reviewed by a Gaelic domain expert.

Finally, the evaluation focuses on linguistic and cultural accuracy rather than general reasoning or mathematical ability, as commonly assessed in English-language benchmarks. Models may exhibit different reasoning performance when operating in a low-resource language setting.

Acknowledgements

This work was carried out by members of the CLARIN Knowledge Centre for Digital Resources for the Languages in Ireland and Britain (DR-LIB) as part of the project 'Unlocking AI for the Languages in Britain and Ireland' project, funded by EPSRC (project number UKRI3181).

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245.

David Adger. 2010. Gaelic morphology. In Moray Watson and Michelle Macleod, editors, *The Edinburgh Companion to the Gaelic Language*, pages 283–303. Edinburgh University Press, Edinburgh.

Muhammad Raheel Anwar, Shah Khalid, Saied Alshahrani, Hafiz Syed Muhammad Bilal, and Mohammed Aldawsari. 2026. MCQs generation with large language models: A survey of methodologies, evolution, and open research issues. *IEEE Access*, 14:10991–11018.

Adnan El Assadi, Isaac Chung, Roman Solomatin, Niklas Muennighoff, and Kenneth Enevoldsen. 2025. Hume: Measuring the human-model performance gap in text embedding tasks.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*,

pages 6265–6281, Albuquerque, New Mexico. Association for Computational Linguistics.

BBC. 2016. Google translate introduces 13 new languages including scots gaelic and sindhi.

BritLLM. 2026. BritLLM: Freely available large language models for UK languages and use-cases. Wayback Machine archive of https://llm.org.uk/. Archived January 21, 2026. Available at: https://web.archive.org/web/20260121124627/https://llm.org.uk/.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Pranjal A Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. 2025. UPDESH: Synthesizing grounded instruction tuning data for 13 indic. *arXiv preprint arXiv:2509.21294*.

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.

Yi R Fung, Chenkai Sun, Jae Doo, Ruining Zhao, and Heng Ji. 2024. No culture left behind: Massively multi-cultural knowledge acquisition & LM benchmarking on 1000+ sub-country regions and 2000+ ethnolinguistic groups. *arxiv*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10.

Pavel Iosad and William Lamb. 2020. Dialect variation in scottish gaelic nominal morphology: A quantitative study. *Glossa*, 5(1):1–31.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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

Josh McGiff, Khanh-Tung Tran, William Mulcahy, Dáibhidh Ó Luinín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O'Sullivan, Hoang D Nguyen, and Nikola S Nikolov. 2025. Irish-BLiMP: a linguistic benchmark for evaluating human and language model performance in a low-resource setting. *arXiv preprint arXiv:2510.20957*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2025. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pages 1907–1917.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale,

Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. 2024. INCLUDE: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: an online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):1–9.

Çağrı Toraman, Ahmet Kaan Sever, Ayse Aysu Cengiz, Elif Ecem Arslan, Görkem Sevinç, Mete Mert Birdal, Yusuf Faruk Güldemir, Ali Buğra Kanburoğlu, Sezen Felekoğlu, Osman Gürlek, et al. 2026. TurkBench: A benchmark for evaluating Turkish large language models. *arXiv preprint arXiv:2601.07020*.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy Chen. 2024. CRAFT: Extracting and tuning cultural instructions from the wild. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 42–47.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Chengxuan Xia, Qianye Wu, Hongbin Guan, Sixuan Tian, Yilun Hao, and Xiaoyu Wu. 2025. Evaluating modern large language models on low-resource and morphologically rich languages: A cross-lingual benchmark across Cantonese, Japanese, and Turkish. *arXiv preprint arXiv:2511.10664*.

Wenda Xu, Sweta Agrawal, Vilém Zouhar, Markus Freitag, and Daniel Deutsch. 2025. Deconstructing self-bias in LLM-generated translation benchmarks. *arXiv preprint arXiv:2509.26600*.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A Hedderich. 2025. MAKIEval: A multilingual automatic Wikidata-based framework for cultural awareness evaluation for LLMs. *arXiv preprint arXiv:2505.21693*.

6. Appendix

6.1. Linguistic Competence Task: System Messages

English You are a knowledgeable assistant that can answer all kinds of questions. Please select the correct option. Output ONLY the letter of the correct option, without any additional explanation or punctuation.

Examples:

> What colour is the sky? ['A. blue', 'B. yellow', 'C. green']. Return ONLY 'A'
>
> Which of these countries is in Africa? ['A. Germany', 'B. Mexico', 'C. Nigeria']. Return ONLY 'C'

Gaelic Is e cuidiche fiosrachail a tha annad agus is urrainn dhut a h-uile seòrsa ceist a fhreagairt. Tagh an romhainn cheart. Na cuir a-mach ach litir na freagairte ceirte, as aonais mìneachadh no pongachadh sam bith eile.

Eisimpleirean:

> Dè 'n dath a th' air an iarmailt? ['A. gorm', 'B. buidhe', 'C. uaine']. Na cuir a-mach ach 'A'
>
> Cò 'n dùthaich às na leanas a tha ann an Afraca? ['A. A' Ghearmailt', 'B. Meagsago', 'C. Nigèiria']. Na cuir a-mach ach 'C'

6.2. Transcript Scoring System Message

You are a Scottish Gaelic article scoring assistant. Given a Gaelic article transcript and its English translation, give a score between 1 and 5 for the cultural relevancy of the article to Gaelic culture. If the article contains no material that relates to Gaelic culture, then give a score of 1. If the article is full of information on important aspects of Gaelic culture, then give a score of 5. For articles somewhere in-between these two, then give a score most fitting the content.

6.3. Cultural Understanding QA system message

You are a Scottish Gaelic question and answer generating assistant. Given a Gaelic article transcript and its English translation, write between 1 and 10 question in Gaelic about the content within the article. The questions should test the answerer's knowledge of Gaelic culture in some way, using only the article as the factual basis for the question and answer. The answers to the questions should

not be easily guessed from the question. Only include as many questions as you are able to make out of the content of the article. Each question should be written so that it makes total sense in isolation and can be answered by someone knowledgeable on the subject without reading the article. Make sure the questions are self contained. You may introduce people, things, places etc. from the article in each question if that helps make the question understandable without reading the article. Do not refer to entities contextually - always use a persons name rather than using 'the man', for example, where possible. Each question can be as long as you would like but should be answerable in less than 10 words. Write the questions and answers in Gaelic and also write an English translation of each.

6.4. Answerability System Message

You are a Scottish Gaelic and English question scoring assistant. Given a question in Gaelic and its English translation, give a score between 1-5 on the self-contained answerability of both questions. Give a score of 5 if the question is a good general knowledge question, is self-contained, is not contextual dependent, and can be answered purely from using knowledge of Gaelic culture. Give a score of 1 if the question is contextual dependent, refers to implicit information outside of general knowledge (e.g. 'the man' rather than 'Robert the Bruce'), or is otherwise badly written. For questions somewhere in-between these two, then give a score most fitting the content. Evaluate the question in both languages on an individual basis, evaluating the wording of the question purely in that language.

6.5. Distractor System Message

You are a Scottish Gaelic distractor generating assistant. Given a Gaelic article transcript and its English translation, as well as a question and answer about the content within the article, write three distractors for the answer. The distractors should be incorrect answers to the question. The distractors should also be plausible while remaining wrong answers. If the correct answer is written in a particular style or format, make sure the distractors also follow this style or format. Remember that there may be multiple correct answers to the original question, so make sure that your three distractors are all completely INCORRECT answers. Write the distractors in Gaelic and also write an English translation for each.