Irish-BLiMP: A Linguistic Benchmark for Evaluating Human and Language Model Performance in a Low-Resource Setting

Josh McGiff[1,*], Khanh-Tung Tran[2,*], William Mulcahy, Dáibhidh Ó Luinín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O'Sullivan[2], Hoang D. Nguyen[2], Nikola S. Nikolov[1]

[1]TEIC Lab, University of Limerick, Ireland
[2]University College Cork, Ireland

Corresponding author: josh.mcgiff@ul.ie

Abstract

We present Irish-BLiMP (Irish Benchmark of Linguistic Minimal Pairs), the first dataset and framework designed for fine-grained evaluation of linguistic competence in the Irish language, an endangered language. Drawing on a variety of linguistic literature and grammar reference works, a team of fluent Irish speakers manually constructed and reviewed 1020 minimal pairs across a taxonomy of 11 linguistic features. We evaluate both existing Large Language Models (LLMs) and fluent human participants on their syntactic knowledge of Irish. Our findings show that humans outperform all models across all linguistic features, achieving 16.6% higher accuracy on average. Moreover, a substantial performance gap of 18.1% persists between open- and closed-source LLMs, with even the strongest model (gpt-5) reaching only 73.5% accuracy compared to 90.1% by human. Interestingly, human participants and models struggle on different aspects of Irish grammar, thus highlighting a difference in representation learned by the models. Overall, Irish-BLiMP provides the first systematic framework for evaluating the grammatical competence of LLMs in Irish and offers a valuable benchmark for advancing research on linguistic understanding in low-resource languages.

Keywords: Linguistic Evaluation, Human Evaluation, Large Language Models, Irish Language

1. Introduction

The rise of the Transformer architecture (Vaswani et al., 2017) has empowered technological advancements across the field of natural language processing (NLP). In terms of NLP, transformers have been instrumental in improving the state–of-the-art for language modelling tasks like sentiment analysis (McGiff and Nikolov, 2024), neural translation (Meta AI, 2024) and dialogue generation (Zhang et al., 2019). The widespread popularity of large language model (LLM) services like ChatGPT and Gemini can be attributed to their capacity to boost productivity and automate complex systems (Filippucci et al., 2024).

Although the impact of transformer-powered services on English and other majority language-based domains is an active research area, low-resource language communities are often an afterthought. Low-resource language communities suffer from the challenges associated with data scarcity and the limited availability of computational resources (McGiff and Nikolov, 2025a). The increasing integration of transformer-powered services in public and private sectors (Department of Public Expenditure, Infrastructure, Public Service Reform and Digitalisation, 2025; QuantumBlack, AI by McKinsey, 2025), despite their limited compatibility with many minority languages, is exacerbating existing forms of technological language inequality. The existing technological biases where English-language content is used by roughly half of the websites on the internet (Common Crawl, 2025), is perpetuated in the training of flagship LLMs. Therefore, more research is required to assess existing LLMs on their understanding of linguistic features in low-resource languages like Irish.

Some work has assessed the linguistic capabilities of natural languages by composing a specialised dataset focused on the evaluation of specific syntactic features. Research has previously focused on majority languages like English (Warstadt et al., 2020) and Chinese (Xiang et al., 2021), with some studies extending to other widely spoken European languages (Suijkerbuijk et al., 2025; Barbini et al., 2025) and non-European languages (Someya and Oseki, 2023). Until now, there has not been any formal research investigating the linguistic capability of LLMs on Celtic languages.

As a result, we have developed Irish-BLiMP (Benchmark of Linguistic Minimal Pairs for Irish). We created 1020 Irish-language minimal pairs for targeted syntactic evaluation. These minimal pairs are created and divided into 11 categories based on linguistic literature relating to the Irish language (Stenson, 2019; Ó hAnluain, 1999; An Caighdeán Oifigiúil, 2017). Some of the 11 macro categories contain sub-categories for finer granularity, with each feature consisting of 10 handcrafted sentences. For example, the verbal morphology category contains various sub-categories that explore each of the verb tenses by standard formation, au-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4575–4586
11–16 May 2026.

tonomous form, irregular formations and other related phenomena. Overall, this paper contributes:

- A novel dataset and an accompanying framework (Irish-BLiMP[1]) for the syntactic evaluation of LLMs on their linguistic understanding of the Irish language.
- A comprehensive experimental analysis revealing a substantial performance gap between state-of-the-art open-source and closed-source models. Open-source models perform near the random baseline, while the strongest closed-source model (gpt-5) achieves only 73.5% accuracy, indicating a significant gap in multilingual grammatical generalisation.
- A human-LLM comparative assessment demonstrating that the benchmark is straightforward for humans (average human accuracy of 90.1%), yet challenging for LLMs.

Irish-BLiMP aims to provide Irish-language LLM researchers with a novel dataset and an accompanying framework for the syntactic evaluation of their models. Unlike typical language model evaluation metrics like BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020), our discriminative evaluation framework assesses existing open-source and proprietary models based on a taxonomy of grammatical features specific to the Irish language. This evaluation framework aims to enable researchers to assess future generative language models based on Irish-language specific features.

2. Related Work

The linguistic evaluation of language models for low-resource languages remains under-researched. Recent studies developing language models for LRLs often rely on automatically calculated metrics like BLEU, COMET and GLUE/SuperGLUE (Wang et al., 2018, 2019; McGiff and Nikolov, 2025a). While these studies highlight a high-level view of the performance of their models, these metrics do not provide granular insights into the syntactic, linguistic understanding of the models. In other words, automatically generated metrics cannot indicate the specific aspects of language that generative models accurately or inaccurately represent. This context is necessary to inform processes for improving language modelling capacity for LRLs.

Previous studies researching language modelling for the Irish language often select from a mixture of evaluation approaches such as using automatically generated metrics and performing human-based evaluation (Barry et al., 2022; Tran et al., 2024a,b; McGiff and Nikolov, 2025b). These approaches suffer from the aforementioned challenge of identifying Irish language-specific linguistic phenomena that the language models struggle with or perform well on. Although some syntactic insights can be found through expert human evaluation of downstream tasks like neural translation (Clifford et al., 2025), a wider assessment of linguistic features is required to highlight and report generalisations across a model's overall understanding of a language like Irish.

An early approach for evaluating the linguistic capabilities of language models introduced the CoLA dataset (Warstadt et al., 2019) by mapping examples from linguistics literature to binary acceptability labels. However, CoLA is limited due to its use of a classifier, and is improved in a subsequent study that introduces BLiMP (Warstadt et al., 2020). BLiMP's approach uses linguist-crafted grammar templates to automatically generate 67 datasets of minimal pairs. A minimal pair consists of an acceptable and an unacceptable sentence that differ only with respect to a single targeted linguistic phenomenon. However, the reliance of BLiMP and other approaches like BLiMP-IT (Barbini et al., 2025) and CLIMP (Xiang et al., 2021) on templates for automatically generating minimal pairs could produce trivial and inaccurate sentences or omit edge-case linguistic features that linguistic experts or literature would cover. Moreover, the frequent use of a limited, predefined vocabulary for automatic generation increases the risk of incomplete coverage of a language model's capacity to capture the full range of syntactic phenomena (Barbini et al., 2025; Xiang et al., 2021). Additionally, given that approaches for morphologically rich languages like Dutch (Suijkerbuijk et al., 2025) opt for a larger number of linguistic feature categories with semi-synthetic minimal pairs, it could indicate that Irish requires more complex templates to create non-trivial minimal pairs in comparison with BLiMP.

JBLiMP (Someya and Oseki, 2023) addresses the pitfalls of automatically generating minimal pairs for Japanese by sourcing 2,323 acceptability judgements from journal articles based on Japanese syntax. Similar to BLiMP, JBLiMP categorised the dataset by type, phenomenon, and paradigm. Similarly, BLiMP-NL (Suijkerbuijk et al., 2025) creates minimal pairs by handcrafting 10 sentence pairs per paradigm. However, BLiMP-NL utilises a language model (GPT-3.5 Turbo) to generate additional minimal pairs from the handcrafted sentence pairs. Although the synthetic sentences were reportedly checked manually and corrected, the generation hinges on the chosen model's linguistic competence for the language as is. For low-resource languages where language models currently have limited understanding of them, this would be more

[1]Code available on GitHub: https://github.com/ReML-AI/Irish-BLiMP

prone to inaccuracies (McGiff and Nikolov, 2025a).

In terms of dataset validation, methods differ between human evaluation using crowd-sourced and expert participants. BLiMP validation participants only validate a fraction of examples per paradigm (Warstadt et al., 2020). In contrast, each minimal pair in JBLiMP is validated by individual participants (Someya and Oseki, 2023). CLiMP enlists at least two native speakers (Xiang et al., 2021) and BLiMP-IT employs at least two linguists (Barbini et al., 2025). Alternatively, BLiMP-NL (Suijkerbuijk et al., 2025) represents the most comprehensive validation in the BLiMP family with each paradigm being tested by at least 30 native Dutch speakers who provided graded acceptability judgements on a 7-point scale and participated in self-paced reading experiments. TurBLiMP, by comparison, only reports acceptability judgement experiments using a 7-point Likert scale with 30 native Turkish speakers (Başar et al., 2025). Overall, studies differ in the number of participants for human validation based on their fluency and expertise in the area of linguistics for the language.

Although a BLiMP-based approach covering 101 languages has been developed (Jumelet et al., 2025), some languages are represented with very few minimal pairs. For example, results are reported for Irish evaluation across various models with 28 minimal pairs. Given that the study only covers basic linguistic phenomena for Irish, the results and insights lack sufficient granularity to meaningfully evaluate model competence in the language. Therefore the MultiBLiMP approach is rather limited in its validity for Irish and other underrepresented languages.

3. Irish-BLiMP

3.1. Data

We compiled a list of acceptability judgements by consulting an array of grammar books for the Irish language (Stenson, 2019; Ó hAnluain, 1999; An Caighdeán Oifigiúil, 2017). Based on these sources, we devised a taxonomy of linguistic phenomena and manually constructed ten minimal pairs for each feature. Our labour-intensive approach of manually creating sentence pairs contrasts with the original BLiMP approach (Warstadt et al., 2020). However, we do not use LLMs to generate minimal pairs.

We enlisted three fluent Irish speakers to produce 1020 minimal pairs spanning 11 linguistic features. Although the final taxonomy includes 11 features aggregated from various linguistic sources, these could be further subdivided to separate closely related phenomena, such as clause structure and word order. Each sentence pair was reviewed by at least two other fluent participants to ensure linguistic correctness. Acceptable and unacceptable sentence pairs were created in alignment with the taxonomy of features provided. It is worth noting that the acceptability judgements are based on the standard form of Irish, known as An Caighdeán Oifigiúil and they do not apply to dialect-specific features (An Caighdeán Oifigiúil, 2017). We report the dataset statistics in Table 1 and describe the data creation process in Figure 1.

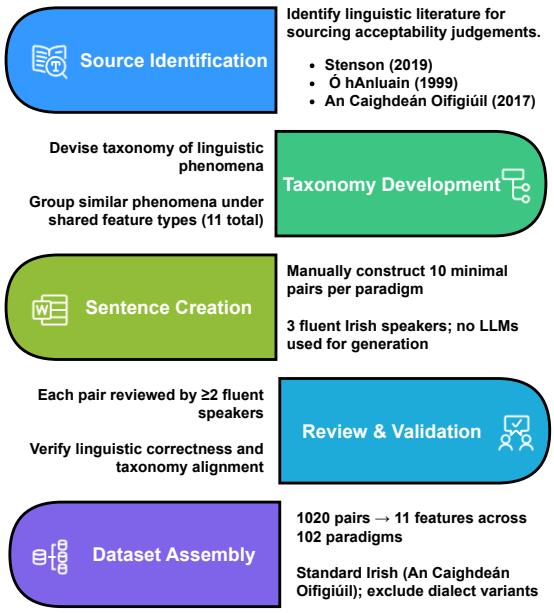

Figure 1: Overview of the data creation process for the Irish minimal pairs dataset, from source identification to final dataset assembly.

3.2. Linguistic Phenomena Coverage

We echo the BLiMP framework (Warstadt et al., 2020) by organising a taxonomy of 11 linguistic phenomena at three hierarchical levels: type, phenomenon and paradigm. Type refers to a broad label encompassing similarly related phenomena. Phenomena denotes specific linguistic features for a given type, such as the case system phenomena as part of the nouns and cases type. A paradigm represents more granular features of a given phenomenon, such as the genitive system as part of the case phenomenon. The following section breaks down the taxonomy by type, phenomena and subsequent paradigms.

3.2.1. Verbal Morphology

This category examines tense, mood, and agreement in verbal forms, including habitual tense contrasts, subject agreement suffixes replacing pro-

Type	Phenomenon	#Paradigms	Total
Verbal morphology	Tense and mood	28	310
	Verb agreement suffixes	2	
	Negation and polarity	1	
Nouns and cases	Gender	1	120
	Case system	5	
	Nouns and adjectives	5	
	Abstract nouns	1	
Adjectives & Comparison	Comparative and superlative degrees	1	40
	Irregular comparatives	1	
	Equative structures	1	
	Adjective predicates with the copula	1	
Pronouns	Personal pronouns	1	90
	Possessive adjectives	2	
	Progressive functions	1	
	Prepositional pronouns	3	
	Relative pronouns	1	
	Pronoun objects	1	
Articles & Determiners	Definite article	1	30
	Article in prepositional contexts	1	
	Determiners	1	
Numbers	Cardinals and nouns	1	30
	Ordinal numbers	1	
	Personal numbers	1	
Copula & bí	Copula uses	6	60
Clause structure & word order	Basic VSO order	1	170
	Other orders	1	
	Progressive sentences	1	
	Verbal nouns	5	
	Subordinate clauses	6	
	Relative clauses	3	
Questions & negations	Yes/No questions	1	50
	Tag questions	1	
	Constituent questions	1	
	Indirect questions	1	
	Answers	1	
Discourse & Sentence types	Reported speech	1	40
	Focus constructions in copula sentences	1	
	Cleft questions and answers	1	
	Embedded clefts	1	
Adverbs & Modifiers	Manner adverbs	4	80
	Time and place adverbs	1	
	Directional adverbs	1	
	Evaluative adjectives	1	
	Sentential adverbs	1	
Total		**102**	**1020**

Table 1: Overview of dataset structure organised by linguistic **Type**, **Phenomenon** and number of **Paradigms**

nouns, and the interaction between negation and emphatic or focus particles.
Example:

- **Acceptable:** *Baineadh geit aisti ar maidin.* ("She got a fright in the morning.")
- **Unacceptable:** *Bhaineadh geit aisti ar maidin.* The autonomous past tense form of a verb is used without lenition.

3.2.2. Nouns and Cases:

This section examines nominal morphology, focusing on gender agreement, case marking across nominative, genitive, dative, and vocative forms, and the interaction of adjectives with case and number distinctions.
Example:

- **Acceptable:** *Cailín bocht.* ("Poor girl.")
- **Unacceptable:** *Cailín bhocht.* Adjectives following nominative singular masculine nouns are not lenited.

3.2.3. Adjectives and Comparison:

This section addresses adjectival morphology and syntax, including comparative and superlative formation, irregular comparison patterns, equative constructions, and the use of adjectives as predicates with the copula.
Example:

- **Acceptable:** *An duine is ciúine.* ("The quietest person.")
- **Unacceptable:** *An duine is ciúin.* The superlative of adjectives ending in a slender consonant is formed by appending *-e*.

3.2.4. Pronouns

This section explores pronominal forms and functions, including personal and possessive pronouns, prepositional and relative constructions, pronoun objects, and the use of pronouns in progressive and possessive contexts.
Example:

- **Acceptable:** *Sin ár n-arán-na.* ("That's our bread.")
- **Unacceptable:** *Sin ár arán-na.* Plural possessive pronouns prepend *n-* to nouns starting with a vowel.

3.2.5. Articles and Determiners

This section examines the use of definite articles, their behavior in prepositional contexts involving mutation, and the role of determiners in expressing definiteness, quantity, and specificity.
Example:

- **Acceptable:** *Ar an lampa.* ("On the lamp.")
- **Unacceptable:** *Ar an nlampa.* *Ar an* causes eclipsis, but *l* cannot be eclipsed.

3.2.6. Numbers

This section addresses interaction of numerals with nouns, including mutation patterns with cardinals, formation and placement of ordinals, and the distinct use of personal numbers for counting people.
Example:

- **Acceptable:** *Ochtar fear.*
  ("Eight men.")
- **Unacceptable:** *Ochtar fhear.*
  Ochtar does not lenite initial consonants.

3.2.7. Copula and Bí

This section examines functions of the copula and the verb *bí*, including their roles in equational and cleft constructions, relative forms, and contrasts in expressing identity, state, possession, and location.
Example:

- **Acceptable:** *Is cluichire mé.*
  ("I am a gamer.")
- **Unacceptable:** *Tá mé cluichire.*
  The copula is required when the predicate is a noun.

3.2.8. Clause Structure and Word Order

This section explores clause organization and syntactic patterns, including basic verb–subject–object (VSO) order, marked word orders for focus or emphasis, the use of verbal nouns in progressive and subordinate constructions, and the structure of complement, causal, temporal, and relative clauses.
Example:

- **Acceptable:** *Labhróidh Liam amárach.*
  ("Liam will speak tomorrow.")
- **Unacceptable:** *Liam labhróidh amárach.*
  The verb labhróidh precedes the subject Liam.

3.2.9. Questions and Negation

This section examines interrogative and negative structures, including yes/no and tag questions, constituent and indirect questions, and the formation of appropriate response sentences.
Example:

- **Acceptable:** *An labhraíonn tú Gaeilge?*
  ("Do you speak Irish?")
- **Unacceptable:** *Ar labhraíonn tú Gaeilge?*
  Present tense questions are formed with the interrogative particle *an*, not *ar*.

3.2.10. Discourse and Sentence Types

This section explores discourse-level and sentence-type variation, including reported speech, copular focus constructions, cleft questions and answers, and embedded clefts within complex clauses.
Example:

- **Acceptable:** *Cheap mé go raibh siad deas.*
  ("I thought they were nice.")
- **Unacceptable:** *Cheap mé go bhí siad deas.*
  The indirect relative clause uses *go raibh*.

3.2.11. Adverbs and Modifiers

This section examines formation and function of adverbs, including manner, time, place, directional types, as well as evaluative and sentential adverbs that convey attitude or discourse-level meaning.
Example:

- **Acceptable:** *Tagann siad anseo go hannamh.*
  ("They come here rarely.")
- **Unacceptable:** *Tagann siad anseo go annamh.*
  An *h* is inserted before a vowel following *go*.

4. Experiment Setup

4.1. Evaluation Method

Our dataset is intended to evaluate the grammatical knowledge of LLMs across diverse linguistic phenomena in the Irish language. We adopt a multiple-choice evaluation paradigm, inspired by the original BLiMP benchmark (which introduces the minimal pair and asks LMs to prefer the grammatical over the ungrammatical option) (Warstadt et al., 2020). This format is widely used and is compatible with frameworks such as lm-evaluation-harness (Gao et al., 2024).

For open-source models, we compute the output probability (log-likelihood) of each choice (A vs. B), then select the option with the higher score as the model's prediction. For API-based (closed-source) models, we prompt the model with the question and compare its textual reply against the correct label. We enforce determinism and reproducibility by setting the temperature to 0, minimise or disable reasoning effort for reasoning-capable models (gpt-5, gpt-oss-120b, and gpt-oss-20b), and require exact-match of the label (A or B).

4.2. Evaluated Models

We benchmark prominent state-of-the-art LLMs. This includes leading closed-source models such as gpt-5 (OpenAI, 2025b), claude-sonnet-4.5 (Anthropic, 2025), and gemini-2.5-flash-lite (Google

DeepMind, 2025), known for superior performance across various tasks, as well as open-source models, namely gpt-oss-120b, gpt-oss-20b (OpenAI, 2025a), Llama-4-Scout-Instruct (Meta, 2025), gemma-3-27b-it (Team et al., 2024), and Mistral-Small-3.2-24B-Instruct (Mistral, 2025).

To our knowledge, none of the frontier multilingual models explicitly support Irish. To this end, we also include UCCIX (Tran et al., 2024b), a model fine-tuned specifically on Irish, as well as its base model (Llama-2-13B (Touvron et al., 2023)), to quantify the effect of Irish-specific adaptation.

4.3. Human Baseline

To ground model performance, we recruited three fluent Irish speakers as independent annotators who were not involved in the dataset creation. Each annotator was prompted with a forced-choice between two sentences via an interactive web application. Each minimal pair prompt was evaluated by each annotator. This yielded a total of 3,060 annotations with an average accuracy of 90.09% (agreement rate: 77.27%), indicating that the minimal pair task is straightforward for fluent speakers and confirming that the items are valid and unambiguous. Moreover, the gap between human and model performance (as shown in the next section) shows that the task remains challenging for current models, prompting further research on LLMs for extremely low-resource languages.

4.4. Prompting Strategies.

We explore three prompting strategies to examine how context or additional examples helps grammatical competence of LLMs:

- Zero-shot: no exemplars or explanations, comparable to the human experiment.
- Few-shot: we create five additional pairs as examples for the LLMs, helping them to understand the task and format requirements.
- Grammar-book context: for each paradigm, a brief description of it is provided alongside the current pair. This helps us analysing in-context grammatical understanding capabilities of LLMs, whether they can learn and deduce the grammatically correct option.

Among these, only zero-shot prompting is directly comparable to the human evaluation setup, as few-shot and grammar-context prompts give LLMs extra advantages not available to humans. Therefore, zero-shot is our baseline that aligns directly with the human test conditions. The prompt templates for each prompting strategy are provided in Appendix A.

5. Results & Discussion

5.1. Performance Gap Between LLMs and Human Baseline

Figure 2 presents the heatmaps of zero-shot prompting accuracy across models and categories. A clear gap emerges between human performance and all evaluated LLMs. Human participants achieved an average accuracy of 90.1%, whereas the top-performing models, gpt-5 and gemini-2.5-flash-lite, reached only 73.5% and 67.1%, respectively. Furthermore, humans consistently outperform all models across all categories, ranking first in all 11 diverse linguistic phenomena.

This trend contrasts with the original English BLiMP benchmark (Warstadt et al., 2020) - intentionally designed to challenge humans and language models alike. In our case, our Irish dataset appears easy for humans but difficult for LLMs, underscoring the disparity in LLM capabilities across extremely low-resource languages. These results highlight the imbalance of multilingual competence and the need for further research on grammar understanding in low-resource linguistic contexts.

5.1.1. Closed-source vs. Open-source Models

Figure 2 also reveals a consistent gap between closed-source (API-only) and open-weight models. gpt-5, claude-sonnet-4.5, and gemini-2.5-flash-lite outperform open models by a wide margin (68.8% compared to 50.7%, averaged across models of the same type), illustrating the performance disparity between proprietary and open models.

Interestingly, gemma-3-27b-it (27B parameters) outperforms models much larger in size (e.g., Llama-4-Scout-17B-16E-Instr. - 109B parameters and gpt-oss-120b - 120B parameters) in overall accuracy. However, this advantage is inconsistent across categories: it underperforms by up to 20.0% on “Adjectives and Comparison” compared to Llama-4-Scout-17B-16E-Instr., but surpasses larger models by over 10 points on “Clause Structure and Word Order”. Nevertheless, all open-source models perform near random chance ($\approx 50.0\%$), indicating that true grammatical understanding of Irish remains limited across current open architectures.

5.1.2. Effects of Language-specific Fine-tuning

Contrary to expectations, language-specific fine-tuning does not lead to significant gains. The Irish-adapted UCCIX model slightly underperforms its base model (Llama-2-13B) by 0.2% average accuracy. This suggests fine-tuning alone may be

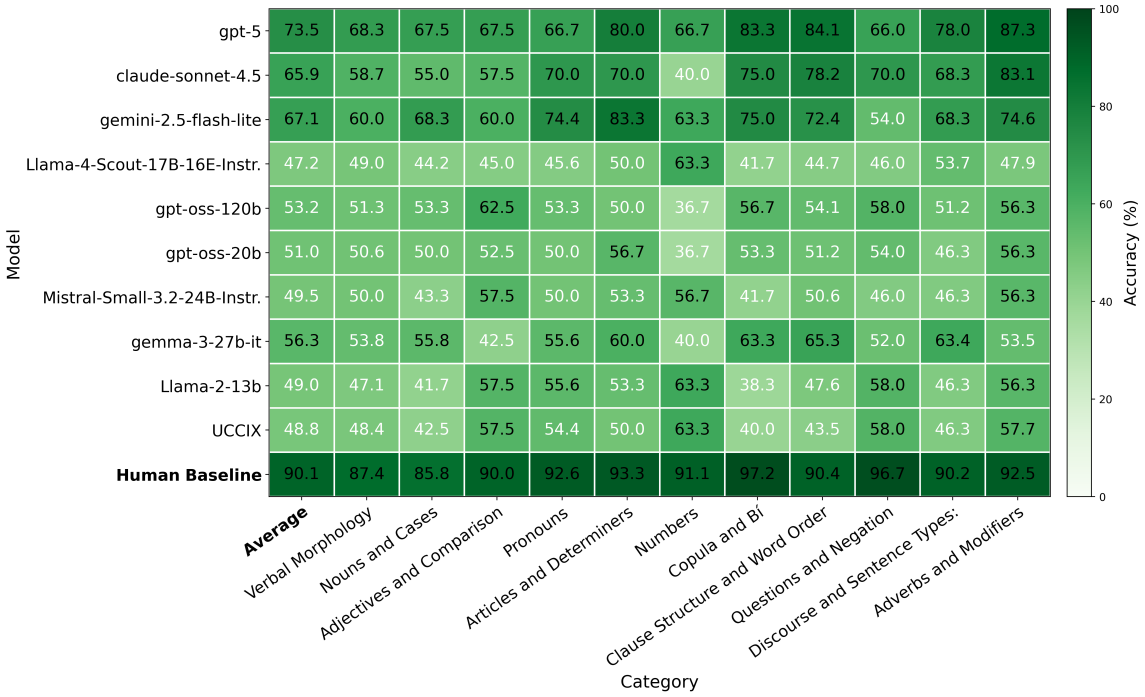

Figure 2: Zero-shot accuracy (%) by model and category.

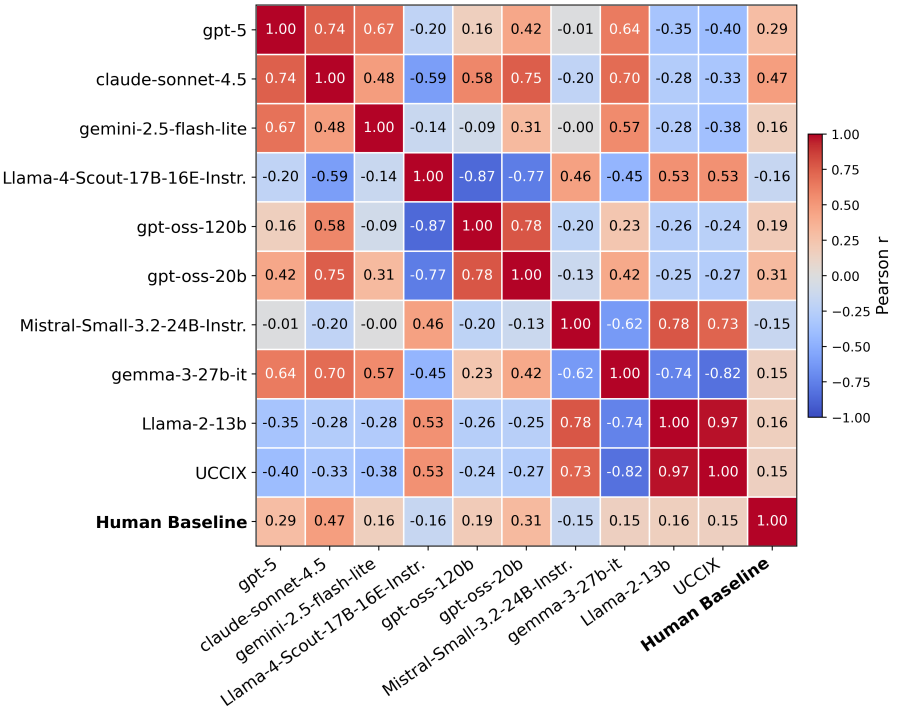

Figure 3: Correlation between models across categories.

insufficient when the available Irish corpus is small, leading to overfitting or limited generalisation, and more efforts are needed for effective multilingual transfer for extremely low-resource languages.

5.2. Performance Variation Across Linguistic Phenomena

We next examine whether certain linguistic phenomena are consistently more challenging for LLMs, if what LLMs found difficult are similar to what humans found difficult. Figure 3 displays the Pearson

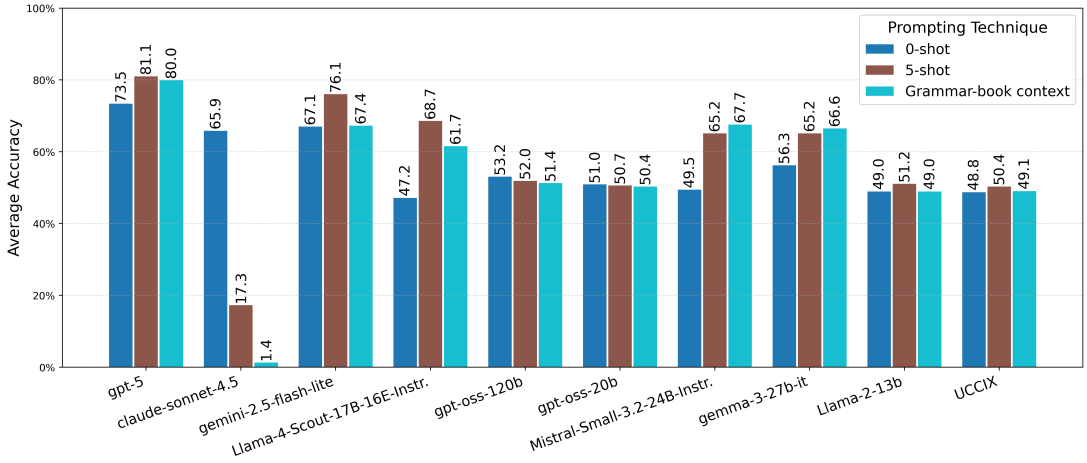

Figure 4: Accuracy by model (groups) and prompting technique (bars).

correlation between human and model accuracies across categories. All LLMs show weak correlation with human performance ($-0.2 < r < 0.5$), with the strongest being claude-sonnet-4.5 ($r = 0.47$). This weak alignment indicates that the types of errors made by models differ from those made by humans, i.e., models struggle with aspects of Irish grammar that humans find trivial.

Moreover, fewer than half of the ten evaluated LLMs exhibit strong inter-model correlation, suggesting that their internal linguistic representations diverge substantially. This reinforces that current LLMs have not yet developed a human-like representation of Irish grammar.

Model	**std.**
gpt-5	8.41
claude-sonnet-4.5	12.24
gemini-2.5-flash-lite	8.55
Llama-4-Scout-17B-16E-Instr.	5.97
gpt-oss-120b	6.51
gpt-oss-20b	5.53
Mistral-Small-3.2-24B-Instr.	5.43
gemma-3-27b-it	8.16
Llama-2-13b	7.69
UCCIX	7.61
Human Baseline	**4.29**

Table 2: Standard deviation across categories.

Table 2 reports the standard deviation of accuracy across phenomena. Human performance shows low variance ($\sigma = 4.29$), implying uniform difficulty. In contrast, LLMs exhibit much higher variance (up to 12.24 for claude-sonnet-4.5), revealing instability in their capabilities across grammatical contexts.

5.3. Prompting Technique Comparison

Figure 4 summarises the accuracy under different prompting strategies. We observe a failure mode for claude-sonnet-4.5 when additional context is provided: instead of producing letter label ("A" or "B") required by the exact-match metric, the model generates extended explanations, leading to failures, potentially due to more attentions are given to the context rather than task requirement.

Overall, both few-shot and grammar-context prompts improve performance for most models, though neither consistently dominates. However, this is not universal, as for some families of model (e.g., gpt-oss and Llama-2), no clear differences are observed over 0-shot prompting. Nevertheless, Llama-4 improves from 47.2% (0-shot) to 68.7% (5-shot), more than 20.0% increase, showcasing strong in-context learning. The best absolute performance is achieved by gpt-5 under 5-shot prompting (81.1%), still below the human baseline (90.1%), even though models receive more contextual information than human evaluators. These results indicate that in-context learning boosts surface accuracy but does not yet yield deep grammatical understanding.

6. Conclusion

In this work, we contribute Irish-BLiMP, a novel dataset for the syntactic evaluation of LLMs and their linguistic understanding of the Irish language, an extremely low-resource language. Our experiments reveal a substantial performance gap: the benchmark is straightforward for humans (average accuracy 90.1%), yet remains highly challenging for LLMs. State-of-the-art open-source models perform near the random baseline, while the strongest closed-source model, gpt-5, achieves only 73.5% accuracy. The performance gap sug-

gests that more work is needed to enable existing models to achieve multilingual grammatical generalisation. To this end, Irish-BLiMP serves as a valuable benchmark for advancing research on linguistic understanding in low-resource languages. Future research could expand coverage to include dialectal variations of Irish.

7. Limitations

This paper focuses specifically on the Irish language, an endangered language. In principle, our framework for creating and evaluating the grammatical competence of LLMs can be extended to other languages as well. Furthermore, our dataset, while covering diverse linguistic phenomena, focuses on written standard Irish, which does not fully capture dialectal diversity. Future work should expand the dataset to include dialect-specific constructions and phonologically conditioned variation.

8. Ethics statement

Our dataset is constructed in alignment with the linguistic features and examples from publicly available sources, ensuring that no copyrighted or sensitive textual material is included. Our work contributes to language technologies that support the digitalisation and preservation of endangered languages, with a particular focus on Irish. We further acknowledge that LLMs trained predominantly on high-resource languages can exhibit systemic linguistic biases. Irish-BLiMP is intended to highlight such disparities and to promote further multilingual model development.

9. Bibliographical References

An Caighdeán Oifigiúil. 2017. *An Caighdeán Oifigiúil: Revised Edition 2017*. Roinn Cultúir, Oidhreachta agus Gaeltachta, Baile Átha Cliath. Department of Culture, Heritage and the Gaeltacht.

Anthropic. 2025. Introducing claude sonnet 4.5.

Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Achille Fusco, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2025. Blimp-it: Harnessing automatic minimal pair generation for italian language model evaluation.

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.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. Turblimp: A turkish benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2506.13487*.

Teresa Clifford, Abigail Walsh, Brian Davis, and Micheál J Ó Meachair. 2025. Gaeilge bhriste ó shamhlacha cliste: How clever are llms when translating irish text? In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 46–51.

Common Crawl. 2025. Common Crawl—Open Repository of Web Crawl Data. Accessed: 2025-07-29.

Department of Public Expenditure, Infrastructure, Public Service Reform and Digitalisation. 2025. Guidelines for the responsible use of ai in the public service. Available at: https://www.gov.ie/en/department-of-public-expenditure-infrastructure-public-service-reform-and-digitalisation/publications/guidelines-for-the-responsible-use-of-ai-in-the-public-service/. Published May 7, 2025; updated May 12, 2025; accessed June 18, 2025.

Francesco Filippucci, Peter Gal, Cecilia Susanna Jona Lasinio, et al. 2024. The impact of artificial intelligence on productivity, distribution and growth.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Google DeepMind. 2025. Gemini flash. https://deepmind.google/models/gemini/flash/. Accessed: 2025-07-28.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Josh McGiff and Nikola S Nikolov. 2025a. Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *arXiv preprint arXiv:2505.04531*.

Josh McGiff and Nikola S. Nikolov. 2025b. Semi-adapt and semilora: Efficient domain adaptation for transformer-based low-resource language translation with a case study on irish.

Josh McGiff and Nikolov S. Nikolov. 2024. Bridging the gap in online hate speech detection: A comparative analysis of bert and traditional models for homophobic content identification on x/twitter. *Applied and Computational Engineering*, 64:63–68.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Meta AI. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Mistral. 2025. Mistral small 3.1.

OpenAI. 2025a. gpt-oss-120b and gpt-oss-20b model card.

OpenAI. 2025b. Introducing gpt-5.

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.

QuantumBlack, AI by McKinsey. 2025. The state of ai: How organizations are rewiring to capture value. Online report, McKinsey & Company. Published March 12, 2025; accessed June 18, 2025.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Taiga Someya and Yohei Oseki. 2023. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594.

Nancy Stenson. 2019. *Modern Irish: A comprehensive grammar*. Routledge.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–35.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024a. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024b. *UCCIX: Irish-eXcellence Large Language Model*. IOS Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Cola: The corpus of linguistic acceptability (with added annotations).

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Liam A. Ó hAnluain. 1999. *Graiméar Gaeilge na mBráithre Críostaí*, eagrán nua edition. An Gúm, Baile Átha Cliath.

A. Prompt Templates

The prompts used for all benchmarked LLMs to generate their responses are shown in Figures 5, 6, and 7, corresponding to the zero-shot, few-shot, and grammar-book context prompting settings, respectively.

```
A: {sentence1}
B: {sentence2}
Answer only the letter of the correct choice:
```

Figure 5: Prompt used in zero-shot prompting strategy.

```
A: {example1 - sentence1}
B: {example1 - sentence2}
Answer only the letter of the correct choice:
{example1 - correct choice}

A: {example2 - sentence1}
B: {example2 - sentence2}
Answer only the letter of the correct choice:
{example2 - correct choice}

A: {example3 - sentence1}
B: {example3 - sentence2}
Answer only the letter of the correct choice:
{example3 - correct choice}

A: {example4 - sentence1}
B: {example4 - sentence2}
Answer only the letter of the correct choice:
{example4 - correct choice}

A: {example5 - sentence1}
B: {example5 - sentence2}
Answer only the letter of the correct choice:
{example5 - correct choice}

A: {sentence1}
B: {sentence2}
Answer only the letter of the correct choice:
```

Figure 6: Prompt used in few-shot prompting strategy.

```
{textbook_section}

A: {sentence1}
B: {sentence2}
Answer only the letter of the correct choice:
```

Figure 7: Prompt used in grammar-book context prompting strategy.