Unveiling the Linguistic Acceptability Judgments of Large Language Models in Multilingual Contexts

Fuyu Xing[1], **Haoyu Huang**[1], **Dawei Mo**[1], **Xinzhuo Yang**[1], **Zixuan Gao**[2], **Wei Wang**[1], **Zimu Wang**[1], **Haiyang Zhang**[1,†]
[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China
[2]Australian National University, Australia
{Fuyu.Xing21,Haoyu.Huang22,Dawei.Mo22}@student.xjtlu.edu.cn
{Xinzhuo.Yang22,Zimu.Wang19}@student.xjtlu.edu.cn
{Wei.Wang03,Haiyang.Zhang}@xjtlu.edu.cn, u7734066@anu.edu.au

Abstract

Linguistic acceptability judgments are essential for evaluating how language models internalize human-like grammatical knowledge. Though some studies have evaluated large language models (LLMs) in this context, existing research lacks systematic exploration of diverse learning paradigms in a multilingual setting. In this paper, we present the first multilingual evaluation of LLMs across four languages (English, Chinese, Japanese, and Russian) in the field of linguistic acceptability. Our evaluation spans both general-purpose (i.e., GPT-4o, GPT-4o mini, DeepSeek-V3, GLM-4-32B, and the Qwen series) and reasoning-oriented (QwQ-32B-Preview and DeepSeek-R1-32B) models under zero-shot and monolingual, cross-lingual and multilingual fine-tuning settings, with comparisons to pre-trained language model (PLM) baselines. Our analysis highlights the strong generalizability of large-scale LLMs through zero-shot prompting, the challenges of fine-tuning small-sized LLMs with skewed training data, the effectiveness of multilingual fine-tuning for low-resource languages, the scaling law exhibited on the task, and the limitation of reasoning-oriented models on the task, even when "aha moments" occur during the reasoning process.

Keywords: Linguistic Acceptability , Multilinguality , Large Language Models

1 Introduction

Measuring the linguistic capability of language models (LMs) is crucial for gaining insights into how they develop human-like linguistic generalizations. A key component of this evaluation involves *linguistic acceptability* judgments, which evaluate the well-formedness and naturalness of sentences from the perspective of native speakers (Fabb, 2019), as examples depicted in Table 1. These judgments are regarded as a fundamental tool in generative linguistics, offering insights into human grammatical knowledge (Chomsky, 1957). Recently, the linguistic competence of LMs has significantly emerged, driven by the advancement of pre-trained models with large-scale architectures, such as pre-trained language models (PLMs, e.g., BERT (Devlin et al., 2019)) and large language models (LLMs, e.g., GPT-4o (Hurst et al., 2024) and DeepSeek (Liu et al., 2025)). These models, pre-trained on vast corpora, have demonstrated superior capability to capture complex linguistic patterns across syntax, semantics, and pragmatics and generate sentences with high grammaticality.

Existing research on the evaluation of linguistic acceptability is still in its infancy. To date, numerous benchmarks have been built for this task, particularly in English (Warstadt et al., 2019; Warstadt et al., 2020), Asian (Hu et al., 2023; Someya et al., 2024), and Indo-European (Trotta et al., 2021; Mikhailov et al., 2022) languages. While these corpora offer valuable insights into models' linguistic abilities, they are predominantly tailored for monolingual evaluations. Therefore, though LLMs have demonstrated promising performance in multilingual, human-like tasks (Chen et al., 2024; He et al., 2024a; Wang et al., 2024a), it remains underexplored whether they can effectively serve as multilingual judges of

†Corresponding author.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 852-863, Jinan, China, August 11-14, 2025.

Label	Sentence
✗	Usually, any lion is majestic.
✓	The men would have been all working.

Table 1: Example linguistic acceptability judgments from the CoLA dataset (Warstadt et al., 2019). ✓ denotes acceptable and ✗ denotes unacceptable.

linguistic acceptability across different evaluation paradigms (Hada et al., 2024; Zhang et al., 2024), highlighting a significant gap in this research field.

In this paper, we present the first comprehensive evaluation to assess the ability of LLMs in multilingual linguistic acceptability. Our evaluation encompasses zero-shot learning as well as fine-tuning in monolingual, cross-lingual, and multilingual settings. Specifically, we select four representative datasets across different languages: CoLA (Warstadt et al., 2019) for English, CoLAC (Hu et al., 2023) for Chinese, JCoLA (Someya et al., 2024) for Japanese, and RuCoLA (Mikhailov et al., 2022) for Russian. In our experiments, we first apply in-context learning (Brown et al., 2020; Peng et al., 2023) on these datasets to evaluate the zero-shot performance of LLMs. We then fine-tune the models using LoRA (Hu et al., 2022) to assess their performance in a more controlled setting. To conduct a comprehensive multilingual evaluation, we further examine the cross-lingual performance by fine-tuning models on English data and evaluating them on the other datasets. Additionally, we assess the multilingual performance to determine whether incorporating more diverse data sources leads to improved results.

We conduct experiments using the following LLMs: GPT-4o (Hurst et al., 2024), GPT-4o mini, DeepSeek-V3 (Liu et al., 2025), GLM-4-32B (Zeng et al., 2024), Qwen2.5-32B (Yang et al., 2025), and Qwen2.5-72B for zero-shot evaluation, and Suzume-Llama-3-8B (Devine, 2024) and Qwen2.5-7B for fine-tuning experiments. Additionally, motivated by the success of reasoning-oriented models, we also include experiments with QwQ-32B-Preview and DeepSeek-R1-32B, both based on the Qwen2.5-32B architecture. Furthermore, we conduct an additional analysis on scaling laws, examining how model performance correlates with model parameters, as well as the impact of reasoning models on this task.

The key findings of this study can be summarized as follows:

- Zero-shot prompting with large-scale LLMs (e.g., GPT-4o and DeepSeek-V3) demonstrates superior performance compared with PLM baselines, particularly in out-of-domain evaluation. It highlights the strong generalizability of LLMs, particularly in tasks such as detecting journal articles and machine-generated content when detecting linguistic acceptability.
- Fine-tuning small-sized LLMs (e.g., Suzume-Llama-3-8B and Qwen2.5-7B) does not consistently outperform PLM baselines across all languages and generally underperforms larger LLMs when directly prompted. These models are also more sensitive to imbalanced label distributions in the training data, and cross-lingual fine-tuning tends to exacerbate label prediction skew.
- Multilingual fine-tuning does not benefit high-resource languages like English and Chinese. However, it can substantially outperform monolingual fine-tuning for low-resource languages such as Japanese and Russian.
- Scaling up model parameters consistently leads to improved performance with some exceptions. Reasoning-oriented models do not exhibit enhanced performance, and incorrect judgments remain unresolved, even after "aha moments."

2 Related Work

Traditional Approaches for Linguistic Acceptability Judgments. Linguistic acceptability judgments have traditionally been formulated as a binary classification task, leveraging PLMs, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), to the exploration in both monolingual settings. Central to this progress are large-scale linguistic acceptability datasets, such as CoLA (Warstadt et al.,

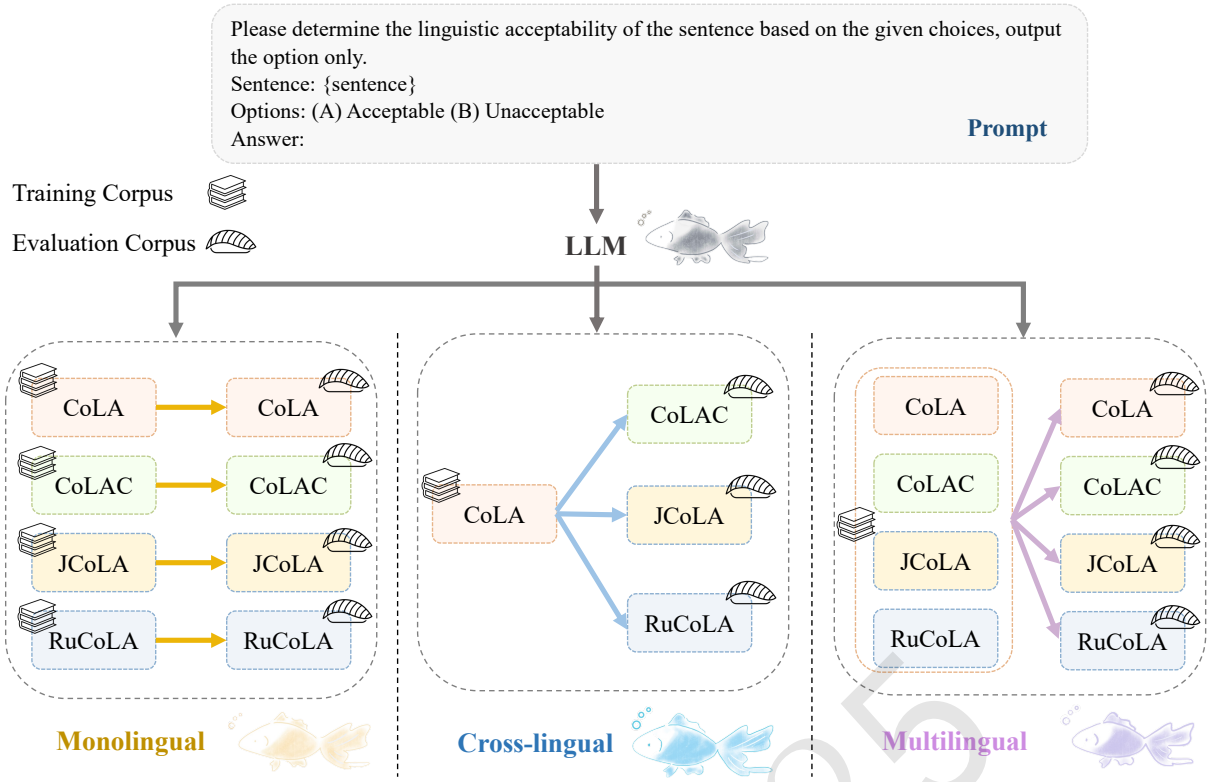

Figure 1: Overall pipeline of the linguistic acceptability evaluation with LLMs, including monolingual, cross-lingual, and multilingual settings. Training corpus is not included in zero-shot evaluations.

2019) for English, CoLAC (Hu et al., 2023) for Chinese, JCoLA (Someya et al., 2024) for Japanese, and RuCoLA (Mikhailov et al., 2022) for Russian, which have provided the foundation for evaluating and improving model performance. This line of work, particularly the adaptation of PLMs through fine-tuning techniques, demonstrates promising potential for addressing the task by integrating linguistic features into the prediction process, equipping the models to capture the nuanced characteristics required for accurate judgment, especially in languages with rich morphological or syntactic variations (Cherniavskii et al., 2022; Proskurina et al., 2023).

LLMs in Linguistic Acceptability Judgments. Motivated by the promising performance of LLMs on various downstream tasks (Qian et al., 2024; Wang et al., 2024b; Ma et al., 2025), they have also been adopted in linguistic acceptability judgments. Vanroy (2024) introduces a German LLM, Feitje, and evaluates the linguistic acceptability performance on the Dutch CoLA dataset. Srinivasan et al. (2024) investigates few-shot tuning and LoRA fine-tuning on OPT models. He et al. (2024b) introduces Consistent Proxy Tuning (CPT), a black-box optimization method to demonstrate improved accuracy on CoLA. However, these efforts remain confined to monolingual settings, and address neither cross-lingual nor multilingual generalizability of LLMs on the task.

Cross-lingual and Multilingual Fine-tuning. In the field of multilingual NLP, methods like InfoXLM (Chi et al., 2021) and Cross-Lingual-Thought Prompting (XLT) (Huang et al., 2023) advance cross-lingual transfer but often overlook linguistic acceptability. In this specific domain, Trotta et al. (2021) pioneers cross-lingual experiments showing that fine-tuning on bilingual data improves performance for Italian acceptability tasks. Hu et al. (2023) further demonstrates acceptability concepts transfer across typologically distinct languages (e.g., English CoLA to Chinese CoLAC). While the MELA benchmark (Zhang et al., 2024) evaluates multilingual linguistic acceptability and establishes zero-shot/few-shot LLM baselines, it overlooks fine-tuned LLM capabilities in cross-lingual and multilingual settings. To bridge this gap, we conduct the first comprehensive evaluation of LLMs with diverse paradigms, providing distinct insights to the development of this field.

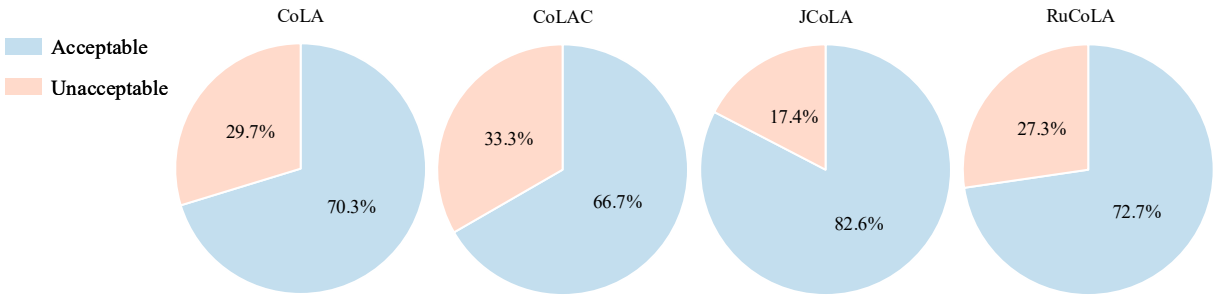

Figure 2: Label distribution of the CoLA, CoLAC, JCoLA, and RuCoLA datasets.

3 Evaluation Settings

Datasets and Evaluation Metrics. We evaluated our approach across four linguistically diverse datasets, each corresponding to a different language. These datasets collectively enable a comprehensive, multilingual assessment of linguistic acceptability judgments. The datasets we utilized are as follows, and the distributions on acceptable and unacceptable samples are organized in Figure 2:

- **Corpus of Linguistic Acceptability (CoLA)** (Warstadt et al., 2019) is used for English evaluation, comprising $8,551$ sentences, with 527 designated for in-domain validation and 516 for out-of-domain validation, each corresponds to different sources and topics.
- **Corpus of Linguistic Acceptability in Chinese (CoLAC)** (Hu et al., 2023) is employed for Chinese evaluation, which includes $6,072$ sentences, of which 492 are reserved for in-domain validation.
- **Japanese Corpus of Linguistic Acceptability (JCoLA)** (Someya et al., 2024) is considered for Japanese evaluation, consisting of $6,919$ sentences, with 865 in-domain and 685 out-of-domain validation instances. The in-domain sources are textbooks and handbooks on Japanese syntax, while out-of-domain sources are journal articles published in JEAL.
- **Russian Corpus of Linguistic Acceptability (RuCoLA)** (Mikhailov et al., 2022) is utilized for Russian evaluation, which contains $7,870$ sentences, including 983 for in-domain and $1,804$ for out-of-domain development. The in-domain sentences are collected from linguistic literature, whereas out-of-domain sentences are produced by machine translation and paraphrase generation models.

For evaluation, we employ two widely used metrics for binary classification tasks: **Accuracy (ACC)** and the **Matthews Correlation Coefficient (MCC)**. These metrics are standard in assessing model performance in acceptability judgment tasks. Their respective formulations are provided below:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{2}$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Prompt Design. Figure 1 presents the prompt design utilized in our evaluation, alongside the procedural framework employed for assessing linguistic acceptability in this study. Generally, the prompt consists of four primary components: an instruction of conducting linguistic acceptability evaluation, an input sentence, a pair of binary options, and a placeholder for the response. We framed the linguistic acceptability task as a multiple-choice question, where LLMs are required to choose between two options—"*Acceptable*" or "*Unacceptable*."

Proceedings of the 24th China National Conference on Computational Linguistics, pages 852-863, Jinan, China, August 11-14, 2025.

Model	CoLA		CoLAC	JCoLA		RuCoLA	
	IN	OUT	DEV	IN	OUT	IN	OUT
Baseline	88.6	82.1	–	**86.4**	82.2	**85.7**	80.1
General-Purpose LLMs							
GPT-4o	85.6	**86.6**	82.7	81.3	**87.0**	81.1	79.5
GPT-4o mini	86.5	85.4	79.3	81.6	84.8	79.7	79.0
DeepSeek-V3	86.9	84.5	**86.4**	78.6	85.8	79.3	**81.9**
GLM-4-32B	85.7	85.2	74.6	85.1	80.4	77.3	73.7
Qwen2.5-32B	87.1	85.4	78.7	82.2	82.2	78.5	77.9
Qwen2.5-72B	**88.8**	85.8	83.1	83.8	85.3	77.2	76.6
Reasoning-Oriented LLMs							
QwQ-32B-Preview	83.8	85.6	78.3	80.0	79.0	76.8	78.1
DeepSeek-R1-32B	85.0	82.9	77.9	81.0	80.6	76.2	77.3

Table 2: Zero-shot evaluation results of LLMs against PLM-based baselines in Accuracy (ACC). The best and the second-best results are in **bold** and underlined, respectively. (IN: in-domain validation, OUT: out-of-domain validation, DEV: validation set of CoLAC)

Experimental Setup. We conducted experiments across a range of LLMs, including both general-purpose and reasoning-oriented models, under zero-shot and fine-tuning settings. For the zero-shot evaluation, we assessed several prominent general-purpose LLMs, such as GPT-4o (`2024-08-06`) (Hurst et al., 2024), GPT-4o mini (`2024-07-18`), DeepSeek-V3 (`0324`) (Liu et al., 2025), GLM-4-32B (`0414`) (Zeng et al., 2024), Qwen2.5-32B (Yang et al., 2025), and Qwen2.5-72B, with temperatures being set as 0 to ensure output stability. Additionally, motivated by the success of reasoning-oriented models in tasks requiring long chains of reasoning, we included experiments with QwQ-32B-Preview[1] and DeepSeek-R1-32B (`Distill-Qwen`) (Guo et al., 2025), both based on the Qwen2.5-32B architecture. This allowed us to investigate whether these models, which excel in providing more granular, step-by-step analysis, could yield improved predictions for linguistic acceptability.

We also fine-tuned several multilingual LLMs, including Suzume-Llama-3-8B (Devine, 2024) and Qwen2.5-7B, using LoRA (Hu et al., 2022) across various settings. First, we performed monolingual fine-tuning on language-specific datasets to evaluate model performance within individual languages. Subsequently, we explored cross-lingual fine-tuning by training on English data and evaluating on other languages to assess knowledge transfer. Finally, we performed multilingual fine-tuning using a combined dataset encompassing all languages to investigate whether incorporating more diverse data sources leads to improved results. During the fine-tuning process, we set the number of epochs to 10, the learning rate to $1e-4$, the batch size to 4, and the gradient accumulation steps to 4. All experiments were conducted on 2 NVIDIA GeForce RTX 3090 graphic cards.

Baselines. We compared the performance of LLMs on linguistic acceptability in comparison to the following state-of-the-art PLM-based models across different languages:

- *English* (CoLA): We employed the fine-tuned BERT model (Devlin et al., 2019) integrated with topological data analysis (TDA) as the baseline, as reported by Cherniavskii et al. (2022).
- *Chinese* (CoLAC): We adopted a direct LLM evaluation approach, as no existing baseline model has been established for the CoLAC dataset.
- *Japanese* (JCoLA): We utilized Waseda RoBERTa[2], a Japanese RoBERTa model pre-trained on Japanese Wikipedia and the Japanese portion of CC-100.

[1] https://huggingface.co/Qwen/QwQ-32B-Preview
[2] https://huggingface.co/nlp-waseda/roberta-large-japanese

Model	CoLA		CoLAC	JCoLA		RuCoLA	
	IN	OUT	DEV	IN	OUT	IN	OUT
Baseline	72.5	56.5	–	**46.6**	50.6	**59.4**	55.8
General-Purpose LLMs							
GPT-4o	66.6	68.2	62.2	40.5	66.0	44.4	54.3
GPT-4o mini	69.1	65.9	54.0	35.5	60.0	39.5	53.4
DeepSeek-V3	71.1	65.4	**71.1**	40.0	**67.5**	41.2	**60.3**
GLM-4-32B	65.3	65.3	44.9	28.9	44.4	27.4	42.1
Qwen2.5-32B	71.6	67.5	53.7	33.6	51.1	34.2	50.7
Qwen2.5-72B	**73.8**	66.9	63.6	36.4	60.0	26.9	47.5
Reasoning-Oriented LLMs							
QwQ-32B-Preview	63.3	**68.7**	51.7	31.0	46.9	32.9	51.0
DeepSeek-R1-32B	65.0	63.0	50.7	29.3	48.4	30.3	49.2

Table 3: Zero-shot evaluation results of LLMs against PLM-based baselines in Matthews Correlation Coefficient (MCC). (IN: in-domain validation, OUT: out-of-domain validation, DEV: validation set of CoLAC)

- *Russian* (RuCoLA): We implemented Ru-RoBERTa (Zmitrovich et al., 2024), enhanced with TDA and two additional features (TDA_{ext}), as reported by Proskurina et al. (2023).

4 Results and Analysis

4.1 Main Results

Zero-shot Evaluation Results. Tables 2 and 3 present the Accuracy (ACC) and Matthews Correlation Coefficient (MCC) scores for the experimented models against baseline methods. From the tables, we observed clear differences in performance between in-domain and out-of-domain settings. For in-domain evaluations, language-specific PLMs consistently outperformed LLMs, achieving superior results in most languages. Notably, these models exceeded the best-performing LLMs by 1.3% in Japanese and 4.6% in Russian. In contrast, LLMs demonstrated a significant advantage in out-of-domain scenarios, showcasing their ability to generalize beyond the training data. For example, GPT-4o outperformed the baseline by 4.5% on English and 4.8% on Japanese datasets, while DeepSeek-V3 surpassed the Russian baseline by 1.8%. These findings highlight the strong generalization capabilities of LLMs, particularly in tasks such as detecting journal articles and machine-generated content when detecting linguistic acceptability.

Our results also highlight the varying performance of LLMs across different languages. While GPT-4o exhibited consistent performance across languages, other models showed notable limitations in specific languages. For instance, GPT-4o mini, GLM-4-32B, and Qwen2.5-32B struggled with Chinese judgments, DeepSeek-V3 underperformed in Japanese judgments, and GLM-4-32B and Qwen2.5-72B did not perform well on Russian judgments, limiting their generalizability across diverse languages. Additionally, we observed that reasoning-oriented models based on Qwen2.5-32B did not outperform the original base model. We will perform a detailed analysis in Section 4.2.

Fine-tuning Evaluation Results. Tables 4 and 5 report the Accuracy (ACC) and Matthews Correlation Coefficient (MCC) results for the fine-tuned models in comparison to baseline methods. Overall, the fine-tuned models did not consistently outperform the baselines across all languages and generally lagged behind larger LLMs when directly prompted. Notably, Qwen2.5-7B outperformed Suzume-Llama-3-8B in all languages except English, likely due to Llama-3's extensive pre-training on large-scale English corpora. In terms of training strategies, multilingual fine-tuning did not benefit high-resource languages such as English and Chinese, leading to performance degradation. However, for low-resource languages like Japanese and Russian, multilingual fine-tuning significantly outperformed monolingual fine-tuning.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 852-863, Jinan, China, August 11-14, 2025.

Model	Setup	CoLA		CoLAC	JCoLA		RuCoLA	
		IN	OUT	DEV	IN	OUT	IN	OUT
Baseline	Monolingual	88.6	82.1	–	86.4	82.2	85.7	80.1
Suzume-Llama-3-8B	Monolingual	**88.8**	**85.1**	**79.5**	82.1	78.1	76.8	72.0
	Cross-lingual	–	–	72.2	81.2	79.3	73.8	70.7
	Multilingual	86.7	84.1	77.9	**83.0**	**79.4**	**77.3**	**74.9**
Qwen2.5-7B	Monolingual	87.5	**84.9**	**82.7**	85.2	**81.5**	80.0	76.1
	Cross-lingual	–	–	75.6	82.1	79.4	75.5	73.2
	Multilingual	**88.4**	**84.9**	82.5	**85.8**	79.4	**81.9**	**77.4**

Table 4: Fine-tuning evaluation results of LLMs against PLM-based baselines in Accuracy (ACC), in which the best result of each model is highlighted in **bold**. (IN: in-domain validation, OUT: out-of-domain validation, DEV: validation set of CoLAC)

Model	Setup	CoLA		CoLAC	JCoLA		RuCoLA	
		IN	OUT	DEV	IN	OUT	IN	OUT
Baseline	Monolingual	72.5	56.5	–	46.6	50.6	59.4	55.8
Suzume-Llama-3-8B	Monolingual	**73.3**	**64.3**	**55.8**	31.0	37.1	35.4	38.9
	Cross-lingual	-	-	38.0	11.7	40.1	11.4	34.3
	Multilingual	68.7	62.0	52.1	**31.1**	**41.8**	**37.0**	**45.4**
Qwen2.5-7B	Monolingual	70.8	63.8	**62.6**	**42.6**	**47.8**	46.3	48.1
	Cross-lingual	–	–	46.0	24.6	41.3	18.7	40.0
	Multilingual	**72.7**	**63.9**	62.1	41.9	40.9	**48.6**	**49.5**

Table 5: Fine-tuning evaluation results of LLMs against PLM-based baselines in Matthews Correlation Coefficient (MCC), in which the best result of each model is highlighted in **bold**. (IN: in-domain validation, OUT: out-of-domain validation, DEV: validation set of CoLAC)

For example, Suzume-Llama-3-8B achieved a 0.5% and 2.9% performance gain on Russian after multilingual fine-tuning, underscoring the value of diverse multilingual data in enhancing performance on under-resourced languages.

On the other hand, the fine-tuned LLMs appeared to be more sensitive to imbalanced label distributions in the training data, as reflected in the substantial drop in MCC scores shown in Table 5—none of the models outperformed the baselines under these conditions. Moreover, the LLMs struggled with cross-lingual transfer, often showing noticeable performance degradation. The most pronounced failure was observed on the Chinese dataset, likely due to the stark structural and grammatical differences between English and Chinese. Furthermore, cross-lingual fine-tuning also contributed to skewed label predictions, further exacerbating the decline in MCC performance.

4.2 Additional Analysis

Scaling Law. Motivated by the performance disparity between Qwen2.5-32B and Qwen2.5-72B, as shown in Tables 2 and 3, we further explored whether scaling up LLMs could enhance performance. Specifically, we selected Qwen2.5 models with sizes ranging from 7B to 72B in a zero-shot learning setup. As depicted in Figure 3, a clear positive scaling trend is evident—larger LLMs consistently outperformed smaller counterparts, with some exceptions, such as the Qwen2.5-14B model on the Japanese dataset. This highlights the importance of leveraging larger models with more advanced language understanding for optimal performance on this task.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 852-863, Jinan, China, August 11-14, 2025.

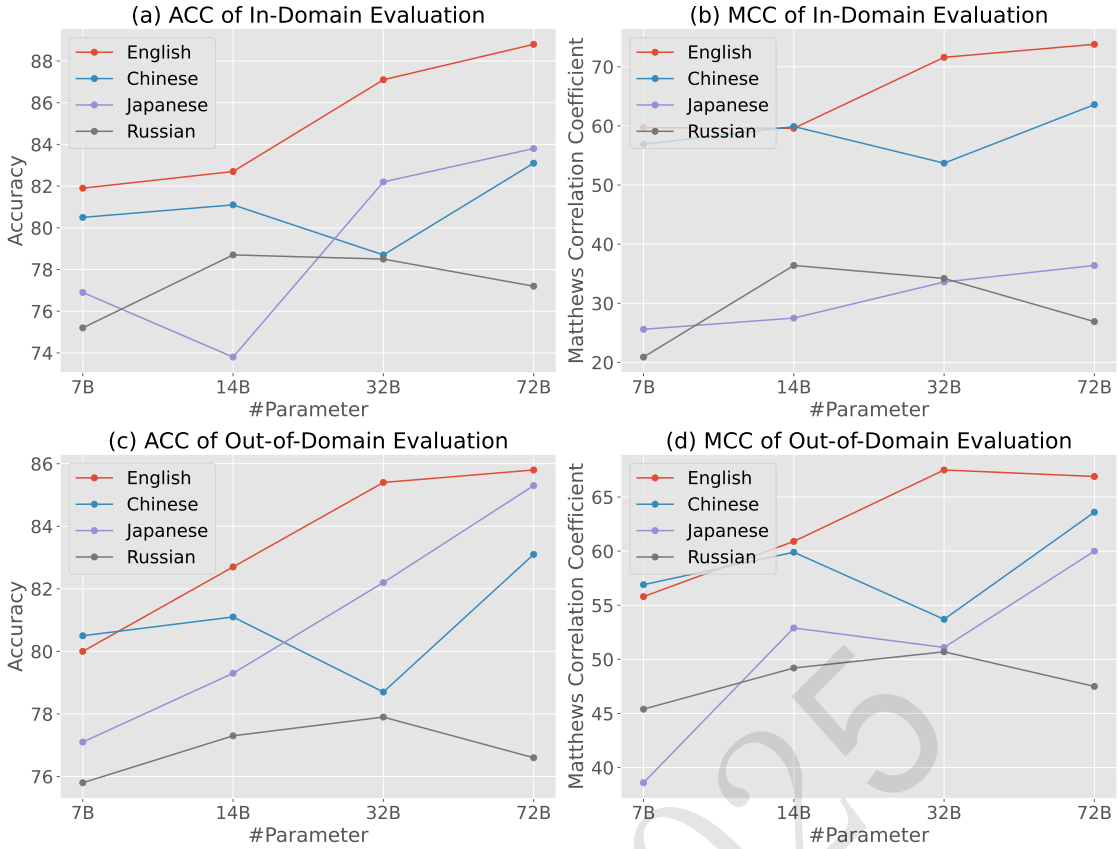

Figure 3: Impact of scaling law on linguistic acceptability judgements, where Qwen2.5-7B/14B/32B/72B are selected for experiments. Results of the CoLAC dataset are illustrated in both in-domain and out-of-domain results.

Impact of Reasoning Chains. As revealed in Tables 2 and 3, it is surprising that reasoning-oriented models exhibited degraded performance. To better understand this phenomenon, we conducted additional case studies to examine how these models perform in the context of linguistic acceptability judgments. In the first example, shown in Table 6, the reasoning model initially understood the task, devised a plan for the judgment process (considering factors like grammar and meaning), and subsequently determines the linguistic acceptability of the sentence based on various aspects (e.g., grammar, sentence structure, punctuation, etc.), ultimately reaching a correct conclusion. However, in the second example shown in Table 7, despite the model generating an extended reasoning chain with several "aha moments"—which are often seen as key indicators of enhanced intelligence in artificial systems (Guo et al., 2025)—the model failed to significantly support the task at hand. It instead maintained its initial, incorrect judgment, underscoring the limitations of reasoning-oriented models in this context.

5 Conclusion and Future Work

We conducted the first comprehensive evaluation of linguistic acceptability using LLMs across diverse languages and learning paradigms, encompassing zero-shot learning as well as fine-tuning in monolingual, cross-lingual, and multilingual settings. In our experiments, we first applied in-context learning to evaluate the zero-shot performance of LLMs. We then fine-tuned the models using LoRA to assess their performance in a more controlled setting. We further examined the cross-lingual performance by fine-tuning models on English data and evaluating them on the other datasets. Furthermore, we assessed the multilingual performance to determine whether incorporating more diverse data sources leads to improved results. Our analysis highlighted the strong generalizability of large-scale LLMs through zero-shot prompting, the challenges of fine-tuning small-sized LLMs with skewed training data, the ef-

TEXT & LABEL: The more pictures of him that appear in the news, the more embarrassed John becomes. (***Acceptable***)
REASONING CHAIN: So I've got this sentence: [...] First, I need to understand what "linguistically acceptable" means. I think it refers to [...] So, I should check the grammar and the meaning of the sentence. Let's look at the **sentence structure**. [...] Now, let's check the **grammar** more closely. [...] Is there any **ambiguity** in the sentence? [...] Are there any **punctuation issues**? [...] What about **word choice**? [...] Is there a better way to **phrase** this sentence? [...] So, overall, I think the sentence is linguistically acceptable.

Table 6: Example reasoning chain of a correctly-predicted sentence. The aspects analyzed by the model is highlighted in **red**.

TEXT & LABEL: Mary listens to the Grateful Dead, she gets depressed. (***Acceptable***)
REASONING CHAIN: But they're connected with a comma. Is that okay? [...] I think this could be a case of a comma splice, where two independent clauses are joined only by a comma without a conjunction. [...] So, maybe that's why it's unacceptable. **Wait, but I've seen sentences like this before,** where two related clauses are connected with just a comma. [...] Let me check some grammar rules. [...] So, according to that rule, the sentence is unacceptable. **But, in informal writing or certain styles,** you might see sentences like this where a comma is used to join two closely related ideas. [...] However, in standard, formal English, it's generally considered incorrect. [...] So, in conclusion, I think the sentence is unacceptable due to the comma splice.

Table 7: Example reasoning chain of an incorrectly-predicted sentence. The "aha moments" within the reasoning chain is highlighted in **red**.

fectiveness of multilingual fine-tuning for low-resource languages, the scaling law exhibited on the task, and the limitation of reasoning-oriented models on the task, even when "aha moments" occur during the reasoning process. In the future, we will expand our evaluation with more languages and LLMs and propose novel methods to improve model performance and validate their utility in practical applications.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This research is supported by the Research Development Funding (RDF) (RDF-21-02-044) and Collaborative Research Project (RDS10120240248) at Xi'an Jiaotong-Liverpool University.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tong Chen, Procheta Sen, Zimu Wang, Zhengyong Jiang, and Jionglong Su. 2024. Knowledge base-enhanced multilingual relation extraction with large language models.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability

judgements via examining the topology of attention maps. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 88–107, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June. Association for Computational Linguistics.

Noam Chomsky. 1957. *Syntactic Structures*. De Gruyter Mouton, Berlin, Boston.

Peter Devine. 2024. Tagengo: A multilingual chat dataset.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Nigel Fabb. 2019. Literature and linguistics, 02.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta, March. Association for Computational Linguistics.

Jianfei He, Lilin Wang, Jiaying Wang, Zhenyu Liu, Hongbin Na, Zimu Wang, Wei Wang, and Qi Chen. 2024a. Guardians of discourse: Evaluating llms on multilingual offensive language detection. In *2024 IEEE Smart World Congress (SWC)*, pages 1603–1608.

Yuanyang He, Zitong Huang, Xinxing Xu, Rick Siow Mong Goh, Salman Khan, Wangmeng Zuo, Yong Liu, and Chun-Mei Feng. 2024b. Cpt: Consistent proxy tuning for black-box optimization.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Patterson, Jiahui Huang, Peng Zhang, Chien-Jer Charles Lin, and Rui Wang. 2023. Revisiting acceptability judgements.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore, December. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2025. Deepseek-v3 technical report.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks.

Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. 2023. Can BERT eat RuCoLA? topological data analysis to explain. In Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybak, Josef Steinberger, and Roman Yangarber, editors, *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 123–137, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Lu Qian, Yuqi Wang, Zimu Wang, Haiyang Zhang, Wei Wang, Ting Yu, and Anh Nguyen. 2024. Domain-specific guided summarization for mental health posts.

Taiga Someya, Yushi Sugimoto, and Yohei Oseki. 2024. JCoLA: Japanese corpus of linguistic acceptability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9477–9488, Torino, Italia, May. ELRA and ICCL.

Krishna Prasad Varadarajan Srinivasan, Prasanth Gumpena, Madhusudhana Yattapu, and Vishal H. Brahmbhatt. 2024. Comparative analysis of different efficient fine tuning methods of large language models (llms) in low-resource setting.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Bram Vanroy. 2024. Fietje: An open, efficient llm for dutch.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In Orphée De Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi, editors, *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand, August. Association for Computational Linguistics.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024b. Generating valid and natural adversarial examples with large language models. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1716–1721.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2025. Qwen2.5 technical report.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: Multilingual evaluation of linguistic acceptability. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand, August. Association for Computational Linguistics.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue,

editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia, May. ELRA and ICCL.