Laughing Across Borders: A Culturally-Aware Joke Generator for Asian Regions

Ashley Fang Cai Xian†, Ng Chen Ting†, Ashley Kok Siu Cheng†,
Wah Yang Tan†, Mohan Raj Chanthran†, Lay-Ki Soon†*, and Meisin Lee†
† Monash University, Malaysia
Email: {afan0007, cngg0048, akok0009, wtan0106}@student.monash.edu,
{mohanraj.chanthran@monash.edu, soon.layki, lee.meisin}@monash.edu

***Abstract*—This paper presents *World Chuckles*, a culturally-aware humour generation system that leverages fine-tuned large language models (LLMs) to generate contextually appropriate jokes tailored to five Asian regions: Malaysia, India, China, South Korea, and Qatar. Addressing the limitations of general-purpose joke generators, the system incorporates a region-specific dataset of annotated jokes and employs fine-tuning techniques to capture local humour styles and cultural nuances. We evaluate the system through a two-phase process: model comparison and user-based evaluation. Participants from each target country assessed the generated jokes on humour quality, cultural relevance, logical structure, and diversity. Results show that the fine-tuned GPT-3.5 Turbo model outperforms LLaMA-3 in producing culturally resonant and engaging jokes. Our work demonstrates the importance of incorporating cultural sensitivity into humour generation and lays the groundwork for future research in personalised, inclusive AI-generated content.**

I. Introduction

According to global estimates, approximately 792 million individuals were living with mental health disorders in 2019, with depression affecting over 264 million people worldwide [1]. With the growing number of mental health cases, it has become increasingly important to develop innovative and accessible tools that can support mental well-being. Personalised jokes tailored to the demographic and interests of the audience may enhance engagement and offer a more meaningful emotional impact. In the current era of Artificial Intelligence (AI), generating jokes has become a relatively simple task using Large Language Models (LLMs), which hold great potential in promoting mental wellness. This technological progress has enabled the rise of online joke generators that aim to offer lighthearted, humour-based content as a means of alleviating emotional distress and enhancing users' mental health.

Despite their potential benefits, current AI-powered joke generators face significant limitations. Based on the latest study by [2], more than 90% of jokes generated by ChatGPT were repeats of a small set of 25 jokes, indicating a reliance on memorised content rather than creating novel jokes. This finding highlights a broader issue in AI-generated jokes which is lack of diversity, originality, and cultural awareness. While several studies have explored humour generation using Large Language Models (LLMs), there remains a critical gap in addressing the cultural and contextual nuances essential to making jokes relatable and respectful.

Language Models used to generate jokes are often built on data that doesn't fully reflect the voices and humour styles of underrepresented groups, especially those from the Global South and minority communities [3]. As a result, their ability to generate jokes is often narrow and biased. These models also have trouble recognising cultural sensitivities, which can lead to jokes that are unintentionally offensive or based on harmful stereotypes, simply because they lack a deep understanding of what's considered appropriate in different cultural contexts. According to the analysis presented in [4], GPT-4o [5] exhibits a significantly higher error rate on jokes related to Chinese culture, with a cultural unawareness error rate of 29.5%, compared to 10.5% for ERNIE Bot [1]. This indicates that GPT-4o is more prone to failure in understanding and generating culturally aware humour, particularly when it involves subtle cultural cues or context-specific references. While both models encounter challenges, GPT-4o demonstrates more difficulty in grasping culturally nuanced humour due to its higher error rate in this area.

To address these limitations, we proposes World Chuckles, a culturally-aware joke generation system that leverages fine-tuned large language models (LLMs) to produce personalised and respectful humour. Specifically, the system focuses on generating jokes that align with the cultural sensibilities of five diverse regions: Malaysia, India, South Korea, China, and Qatar. By incorporating regional language patterns, local references, and cultural norms into the training process, the proposed model aims to move beyond generic joke templates and instead deliver content that feels more relatable and engaging to users from different backgrounds. The contribution of the work can be summarised below:

1) Construction of Dataset: Developing a multi-regional dataset incorporating culturally relevant jokes from five target countries, supporting more inclusive and representative humour generation.
2) Cultural-Aware Joke Generation Model: Development of a culturally aware joke generation model that integrates fine-tuning techniques and region-specific data to reflect

*Corresponding Author.

[1]https://research.baidu.com/Blog/index-view?id=183

979-8-3315-7206-8/25/$31.00

diverse humour styles.

3) Contextual Humour Evaluation Framework: Design of an evaluation framework to assess joke quality, cultural relevance, and user perception across different cultures, highlighting the impact of cultural alignment in AI-generated humour.

This paper is structured as follows: Section II reviews existing approaches to joke generation models. Section III introduces the Cultural-Aware Joke Generation model and discusses the methodology in detail. Section IV and Section V describe the experimental setup and present the results, respectively. Section VI concludes the paper and outlines directions for future work. Finally, Section VI provide acknowledgements.

II. Related Works

A. Joke Generation Using Classification Approaches

Early joke generation systems often relied on classification techniques. One of the earliest approaches is called Generalised Analogy Generator (GAG), which uses Random Forest classifiers and joke rating templates to generate new jokes [6]. Although this approach identified important humour features, it failed to consider rating order and lacked the ability to model humour nuance, resulting in low-quality joke outputs.

Another work by Yamane et al. [7] proposed a morally-aware joke generator using a predefined template ("I like my X like I like my Y, Z") and a moral classifier. While it successfully filtered out offensive content, the rigid format limited joke diversity, and the model showed no significant improvement in perceived funniness.

B. Joke Generation via Fine-Tuned Language Models

Recent advancements have shifted towards fine-tuning large language models (LLMs) such as GPT-2 and BERT. Akbar et al. [8] fine-tuned GPT-2 for joke generation and BERT for classification using the Short Jokes dataset and non-joke data. The model effectively captured the setup–punchline structure and outperformed earlier baselines, but still struggled with generalising humour across different contexts and was prone to being tricked by simple patterns.

Efforts to expand humour detection into multilingual contexts have also shown progress. Weller and Seppi [9] fine-tuned BERT on humour-specific corpora, outperforming CNN models, while Ismailov [10] enhanced performance on the IberLEF 2019 HAHA dataset by combining multilingual BERT with traditional classifiers such as Naive Bayes and logistic regression. Despite these gains, such models remain computationally expensive and often lack the ability to generalise across underrepresented humour styles.

Most existing models focus on English and Western humour, neglecting the cultural, linguistic, and ethical dimensions crucial to joke relevance and appropriateness. Current approaches also fall short in generating humour that is both diverse and respectful across different regions.

III. Methodology

To address the limitations of existing methods, our proposed system, World Chuckles, focuses on fine-tuning large language models (LLMs) to generate culturally-aware jokes. By using region-specific datasets from Malaysia, India, South Korea, China, and Qatar, the model learns local humour styles and produces jokes that are both engaging and culturally appropriate. As shown in Fig 1, the system follows a structured workflow consisting of four main stages: Data Collection and Annotation (which is discussed in Section III-A), and Model Fine-Tuning (which is discussed in Section III-B). Country-specific jokes were first collected from online sources, then annotated by recruited participants who rated their humour quality and cultural sensitivity (offensive vs. non-offensive). These annotated jokes were used to fine-tune LLMs (GPT-3.5 Turbo and LLaMA), enabling the models to capture the cultural nuances of humour across different regions.

A. Data Collection and Annotation

To train the model on culturally relevant humour, a total of 500 jokes were collected from various online sources such as social media, websites, YouTube, and articles. Each joke was labelled with the corresponding country. For the annotation process, 25 participants from different backgrounds were recruited to rate the collected jokes. We have a total of 25 participants, with jokes sourced from five regions: Malaysia, India, China, South Korea, and Qatar. Each region is represented by 5 participants. All participants are students from Monash University, with a diverse mix of genders. The ratings included:

1) Cultural Sensitivity: Classified as Offensive or Non-Offensive.
2) Quality: Rated on a 1–5 star scale based on humour and relevance

The dataset included 100 jokes per country, with approximately 30% offensive and 70% non-offensive content to help the model learn boundaries of cultural sensitivity. Each participant rated 20 unique jokes specific to their country of origin, ensuring feedback aligned with local cultural understanding.

B. Model Selection and Fine-Tuning

In the model training phase, we fine-tuned two large language models, Llama-3 [11] and GPT-3.5 Turbo [12], using the preprocessed and rated joke dataset. These models were intentionally selected to represent two prominent families of large language models: Llama 3, as a leading open-source model, and GPT-3.5 Turbo, as a widely used proprietary model accessed via API. This distinction allowed us to explore how models with different architectures and access paradigms perform in humor generation. Preprocessing involved cleaning the joke text, encoding country tags, and formatting inputs for compatibility. Fine-tuning leverages transfer learning to adapt the pre-trained models to the culturally annotated humour domain. The goal was to evaluate which model performed better in generating culturally sensitive, logical, and humorous jokes.

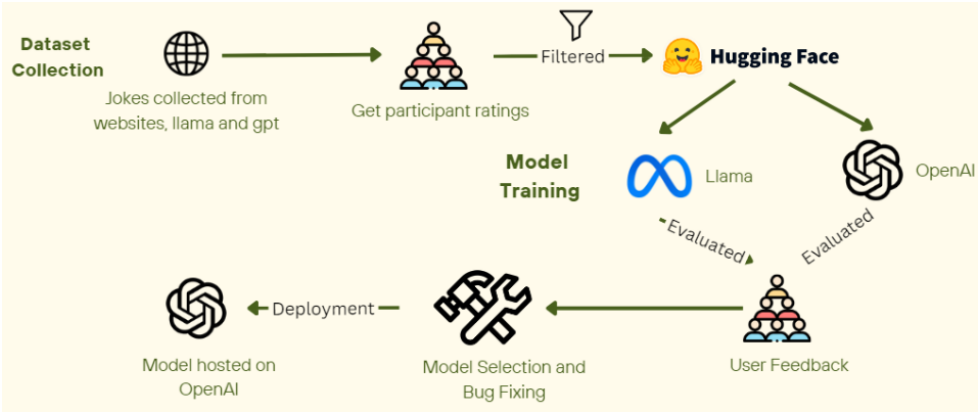

Fig. 1. End-to-end process of developing World Chuckles system. The process includes data collection of country-specific jokes, annotation through user-based cultural sensitivity and quality ratings, fine-tuning of large language models (LLMs) using the annotated dataset, and final evaluation through user testing to assess cultural relevance, humour quality, and offensiveness.

C. Fine-tuning Configuration

For fine-tuning, we employed the `SFTTrainer` from the `trl` [2] library with a causal language modeling objective. Training was conducted over **5 epochs** to ensure sufficient learning without overfitting. A batch size of **4** per device was chosen to balance memory constraints and training stability, while a learning rate of $\mathbf{3 \times 10^{-5}}$ was selected based on common best practices for fine-tuning large language models to facilitate steady convergence.

We used the `paged_adamw_32bit` optimizer to efficiently handle model parameters with memory optimization and applied a weight decay of **0.001** to regularize and prevent overfitting. Gradient clipping with a maximum norm of **0.1** was applied to maintain stable training dynamics, and a warm-up ratio of **3%** was used to gradually ramp up the learning rate, avoiding early training instability. Grouping sequences by length improved training efficiency by minimizing padding overhead. Logging and checkpointing every **25 steps** allowed for frequent monitoring without excessive interruption, with `TensorBoard` used for visualization.

To reduce the number of trainable parameters and improve training efficiency, we applied LoRA-based parameter-efficient fine-tuning with a rank $r = 64$, a LoRA alpha of **16**, and a dropout rate of **0.1**. The parameters used were selected based on results from various training iterations.

IV. Evaluation

To assess the effectiveness of our culturally-aware joke generation system, we conducted a Comparative Model Evaluation. The evaluation aimed to test how well the models performed and how people reacted to the jokes they produced. A total of 25 participants, who took part during our data collection phase, were asked to review jokes generated by Llama-3 and GPT-3.5 Turbo using a structured evaluation form. Participants are required to review jokes based on four main aspects: how relevant the joke was to the culture, how funny it was, whether it made logical sense, and how different or unique the jokes were. Each participant rated the jokes using a 1 to 5 scale, where 1 means "Very Poor" and 5 means "Excellent". This helped us understand which model was better at generating culturally appropriate and enjoyable jokes.

[2]https://github.com/huggingface/trl

Cultural Region	Model	Number of Ratings per Likert Scale Value (1-5)				
		1	2	3	4	5
Malaysia	Fine-tuned Llama-3	1	10	11	3	0
	Fine-tuned GPT-3.5 Turbo	0	2	2	16	5
India	Fine-tuned Llama-3	0	3	10	7	5
	Fine-tuned GPT-3.5 Turbo	0	1	4	14	6
South Korea	Fine-tuned Llama-3	2	4	10	8	1
	Fine-tuned GPT-3.5 Turbo	1	4	5	10	5
China	Fine-tuned Llama-3	2	2	17	3	1
	Fine-tuned GPT-3.5 Turbo	0	4	6	11	4
Qatar	Fine-tuned Llama-3	0	15	6	2	2
	Fine-tuned GPT-3.5 Turbo	1	7	12	3	2

TABLE I
User Evaluation Ratings of Generated Jokes Across Different Cultural Regions

V. Result and Discussions

To evaluate the effectiveness of our joke generation models, we conducted a comparative assessment between the fine-tuned LLamA-3 and GPT-3.5 Turbo models. As discussed in Section IV, this evaluation aimed to determine which model better captures cultural nuances and produces higher-quality, culturally-aware jokes for our target countries: Malaysia, China, South Korea, Qatar, and India. Table I shows the detailed Likert-Scale

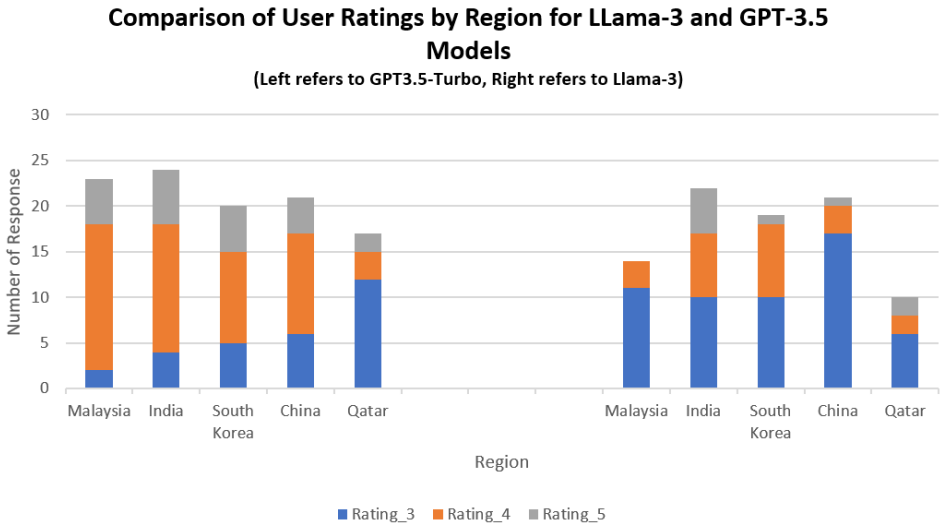

Fig. 2. Side-by-side stacked bar chart comparing the number of user ratings (Rating 3, 4, 5) for jokes generated by the fine-tuned Llama-3 and GPT-3.5 Turbo models across five regions.

ratings by country, providing an aggregated view of participant responses. Each row represents a specific country, while the columns reflect the distribution of scores from 1 (Very Poor) to 5 (Excellent).

Fig 2 presents a comparative bar chart illustrating the distribution of user ratings for both models. The fine-tuned GPT-3.5 Turbo model received a notably higher number of 4-star and 5-star ratings, particularly in terms of humour quality and cultural relevance. In contrast, the LLamA-3 model saw a concentration of 3-star ratings, indicating relatively average performance. These findings suggest that GPT-3.5 Turbo consistently produced more engaging and culturally appropriate jokes across all five countries. Based on the outcomes of this comparative evaluation, we selected the fine-tuned GPT-3.5 Turbo model for integration into the World Chuckles web application. This decision was supported by the model's consistent ability to generate jokes that were better aligned with cultural expectations and user preferences.

From a broader perspective, these findings demonstrate the value of fine-tuning large language models with culturally annotated datasets. The evaluation confirms that cultural sensitivity in AI-generated content is not only feasible but also essential for improving engagement and user satisfaction. Existing humour generation tools often overlook this dimension, resulting in jokes that may be perceived as irrelevant or even offensive. In contrast, World Chuckles leverages cultural context to produce humour that resonates with users, ultimately contributing to more inclusive and respectful AI applications.

VI. Conclusions

In conclusion, this study introduced World Chuckles, a culturally-aware joke generation system that leverages fine-tuned LLM to produce contextually appropriate humour across five Asian regions—Malaysia, India, South Korea, China, and Qatar. By fine-tuning both Llama-3 and GPT-3.5 Turbo models on region-specific datasets and conducting structured user evaluations, the GPT-3.5 Turbo model was selected for deployment due to its superior performance in generating jokes that were culturally relevant, humorous, and logically coherent. This work highlights the importance of cultural sensitivity in humour generation and demonstrates the effectiveness of user feedback in refining AI-generated content. For future work, we aim to expand the dataset to include more regions and languages, integrate multilingual capabilities, explore dynamic prompt adaptation based on user preferences, and implement real-time feedback loops using reinforcement learning to further personalise and enhance joke quality.

Acknowledgment

We would like to acknowledge the responsible use of Generative AI tools, which assisted in error checking and improving the clarity of my writing in compliance with academic integrity guidelines. Large Language Models (LLMs) like Llama-3 and GPT-3.5 Turbo were used for fine-tuning purpose.

References

[1] S. Dattani, L. Rodés-Guirao, H. Ritchie, and M. Roser, "Mental health," *Our World in Data*, 2023, https://ourworldindata.org/mental-health.

[2] S. Jentzsch and K. Kersting, "ChatGPT is fun, but it is not funny! humor is still challenging large language models," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, J. Barnes, O. De Clercq, and R. Klinger, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 325–340. DOI: 10.18653/v1/2023.wassa-1.29. [Online]. Available: https://aclanthology.org/2023.wassa-1.29/.

[3] P. Mirowski, J. Love, K. Mathewson, and S. Mohamed, "A robot walks into a bar: Can language models serve as creativity supporttools for comedy? an evaluation of llms' humour alignment with comedians," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24, Rio de Janeiro, Brazil: Association for Computing Machinery, 2024, pp. 1622–1636, ISBN: 9798400704505. DOI: 10.1145/3630106.3658993. [Online]. Available: https://doi.org/10.1145/3630106.3658993.

[4] R. He, Y. He, L. Bai, *et al.*, "Chumor 1.0: A truly funny and challenging chinese humor understanding dataset from ruo zhi ba," *arXiv preprint arXiv:2406.12754*, 2024.

[5] OpenAI, : A. Hurst, *et al.*, *Gpt-4o system card*, 2024. arXiv: 2410.21276 [cs.CL].

[6] T. Winters, V. Nys, and D. De Schreye, "Automatic joke generation: Learning humor from examples," in *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts: 6th International Conference, DAPI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II*, Las Vegas, NV, USA: Springer-Verlag, 2018, pp. 360–377, ISBN: 978-3-319-91130-4. DOI: 10.1007/978-3-319-91131-1_28. [Online]. Available: https://doi.org/10.1007/978-3-319-91131-1_28.

[7] H. Yamane, Y. Mori, and T. Harada, "Humor meets morality: Joke generation based on moral judgement," *Information Processing Management*, vol. 58, no. 3, p. 102 520, 2021, ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102520. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321000297.

[8] N. Akbar, I. Darmayanti, S. Fati, and A. Muneer, "Deep learning of a pre-trained language model's joke classifier using gpt-2," *Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences*, vol. 48, p. 2021, Jul. 2021.

[9] O. Weller and K. Seppi, "Humor detection: A transformer gets the last laugh," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3621–3625. DOI: 10.18653/v1/D19-1372. [Online]. Available: https://aclanthology.org/D19-1372/.

[10] A. Ismailov, "Humor analysis based on human annotation challenge at iberlef 2019: First-place solution.," 2019. [Online]. Available: https://ceur-ws.org/Vol-2421/HAHA_paper_3.pdf.

[11] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[12] OpenAI, *GPT-3.5 Turbo*, https://platform.openai.com/docs/models/gpt-3-5, Accessed: 2025-07-04, 2023.