Thai Humor Generation by Small Language Models

Panjapong Poobanchean
Computer Engineering Department
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
panjapong.pooba@kmutt.ac.th

Sorawit Chaithong
Computer Engineering Department
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
sorawit.chait@kmutt.ac.th

Charunthon Limseelo
Computer Engineering Department
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
charunthon.lims@kmutt.ac.th

Kantapat Suwannahong
Computer Engineering Department
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
kantapat.suwan@kmutt.ac.th

Nuttanart Muansuwan
Computer Engineering Department
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
nuttanart.mua@mail.kmutt.ac.th

Booncharoen Sirinaovakul
Computer Engineering Department
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
boon@kmutt.ac.th

***Abstract*— Despite the impressive capabilities of generative AI across multiple languages, generating humor that aligns with Thai cultural and linguistic nuances remains a significant challenge. Thai humor often relies on context, wordplay, and socio-cultural references, making it difficult for generic models to produce authentic jokes. This paper presents a focused approach to address this limitation by fine-tuning small language models (SLMs) on high-quality, non-synthetic Thai humor datasets. Llama-3.2-3B model was leveraged and Low-Rank Adaptation (LoRA) was employed for efficient parameter tuning, ensuring computational efficiency suitable for low-resource settings. Our work highlights humor as a critical benchmark for evaluating AI's understanding of language semantics and cultural context. A comprehensive evaluation was conducted with Thai participants to ensure the generated humor resonates with real-world cultural expectations.**

Keywords*— *Generative AI, Thai humor, small language models (SLMs), fine-tuning, non-synthetic datasets, Low-Rank Adaptation (LoRA)

I. Introduction

The development of progressively large language models (LLMs) has showcased impressive natural language generation abilities, especially in collaborative tasks with humans such as technical and creative writing. Leveraging LLMs for research and potential uses in higher education, however, poses challenges in terms of accessibility and adaptability of infrastructure resources. Transferring knowledge from LLMs to more accessible Small Language Models (SLMs) can potentially be a solution. This research explores the generation of Thai humor using small language models (SLMs), focusing on how artificial intelligence (AI) can create humorous contents within specific linguistic and cultural contexts of Thailand. The study will take into consideration the types of humor, linguistic patterns, structures, and cultural elements that are integral to Thai humor, with the goal of developing effective methods for fine-tuning SLMs to produce humorous contents.

II. Literature Review

In this section, theories of humor and AI's capacity for humor generation are explored in different aspects.

A. Theories of Humor

Traditional theories of humor are relating to many ways humans feel such as the feelings of superiority, relief and incongruity. Laughter can be generated from feeling superior, or releasing of some negative energy such as nervousness. Incongruity is also the cause of laughter as it happens due to unexpected situations [1], [2]. Moreover, according to [1], humor arises as a result of violating norms or expectations that are simultaneously perceived as harmless. This is particularly crucial for understanding humor in memes, which often rely on playful subversions of cultural norms. Additionally, individual difference is a factor affecting the appreciation of humor. The difference is influenced by individual personality traits, cultural background, and emotional states and it needs to be accounted for as individual variations in humor responses [2].

B. AI and Humor Generation

Advances in generative AI, particularly deep neural networks, have enabled the creation of high-quality images and texts with tools such as ChatGPT, DALL-E, and Midjourney [1]. However, generating humor still presents unique challenges for AI as it requires understanding cultural contexts, nuanced language, and the subtle interplay of expectations and violations. Moreover, research suggests that people may be less receptive to AI-generated humor due to a phenomenon called "algorithm aversion"[3]. This bias against algorithmic outputs can influence how people perceive and evaluate AI-generated jokes. Other factors to consider based on several empirical findings include beliefs about AI creative limitations [4] and originality [5] which result in rating AI-generated jokes as less funny or not original because it lacks comedic timing, for example. Finally, people are not always looking for humorous responses but personalized ones such as in chatbot recommendations, particularly for travel recommendations [6]. This suggests that different humor styles might be more effective in certain contexts.

III. METHODOLOGY

A. Data Collection, Library and Environment Setup

The joke data were collected from comedy clips, comedy shows and websites for 1,800 rows dataset.

TABLE I. DATASETS DESCRIPTION

Number of samples	***Source***
58	YouTube Clip
15	Comedy shows
1,674	Websites

The research was conducted using the Google Colab environment. The primary libraries used in the study include Unsloth, a parameter-efficient fine-tuning library for LLMs, and Hugging Face Transformers, which provides pre-trained models, tokenizers, and tools for natural language processing tasks. Hugging Face Transformer Reinforcement Learning (TRL), built on Transformers, offers functionalities for reinforcement learning and supervised fine-tuning of LLMs. Datasets is a library for loading and processing datasets in various formats, while Torch is a deep learning framework for building and training neural networks. Matplotlib is a plotting library for creating visualizations of data and model outputs, and Numpy is a library for numerical computing in Python. These libraries provide the foundation for data loading, model fine-tuning, inference, and evaluation within the research workflow.

B. Model Selection and Finetuning

The research uses Meta's Llama-3.2-3B-Instruct model [7] as the base model for fine-tuning tasks. Fine-tuning is performed using the Hugging Face TRL library's SFTTrainer, which offers a high-level API for fine-tuning LLMs on custom datasets. To improve efficiency, LoRA (Low-Rank Adaptation) [8] is employed, which injects trainable rank decomposition matrices into each layer of the Transformer model, reducing the number of trainable parameters.

The training parameters are carefully configured to optimize the model's performance. These parameters include:

- per_device_train_batch_size: The number of training samples processed per device in each training step, set to 2 in this study.
- gradient_accumulation_steps: The number of steps over which gradients are accumulated before performing a weight update, set to 4 to effectively increase the batch size without exceeding memory limits.
- warmup_steps: The number of initial training steps where the learning rate is gradually increased, set to 5 to prevent large initial updates that could destabilize the training.
- max_steps: The maximum number of training steps, set to 60 for demonstration purposes; for a full training run, num_train_epochs can be set to 1 and max_steps to None.
- learning_rate: The initial learning rate for the optimizer, set to 2e-4.
- fp16 or bf16: The floating-point precision used during training, automatically determined based on hardware capabilities to optimize speed and memory usage.
- logging_steps: The frequency of logging training progress, set to 1 to monitor performance closely.
- optim: The optimizer used for updating model weights, set to "adamw_8bit" for efficient optimization.
- weight_decay: A regularization term to prevent overfitting, set to 0.01.
- lr_scheduler_type: The learning rate scheduling strategy, set to "linear" for a gradual decrease in learning rate over time.
- seed: The random seed used for reproducibility, set to 3407.
- output_dir: The directory where training outputs are saved, set to "outputs".
- report_to: The platform for reporting training metrics; set to "none" in this case.

C. Inference and Evaluation

After completing the fine-tuning process, the model is ready for inference, which involves generating text based on input. The FastLanguageModel.for_inference method optimizes inference speed by enabling faster native execution paths. The model's performance is evaluated qualitatively by providing Thai joke prompts, which are assessed by human evaluators based on humor, coherence, and relevance to the prompt. Humor refers to elements that evoke laughter or amusement, coherence assesses logical flow and grammatical correctness, and relevance determines if the joke aligns with the prompt and the intended topic.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Prompting Methods

Natural language generation based on LLMs can perform new tasks it was not trained for, in this case to generate jokes, by finetuning using Zero-Shot or Few-Shot prompting. In this research, we experimented with how this process works and see the performance of prompting methods for generating new jokes, which will be further evaluated on how funny they are.

Jokes were generated using three prompting methods: Zero-Shot Prompting, One-Shot Prompting, and Few-Shot Prompting. Each method offered a unique approach to guiding the model's response, varying in the level of context provided. Below is a table that demonstrates the structure of the prompts used in each prompting method, followed by an example.

TABLE II. PROMPTING METHODS AND EXAMPLES

Prompting Method	***Prompt Structure***	***Prompt Example***
Zero-Shot	“คุณคือตลกที่เก่งที่สุดในประเทศไทย จงคิดมุกตลกที่เกี่ยวกับ […X…]”	“คุณคือตลกที่เก่งที่สุดในประเทศไทย จงคิดมุกตลกที่เกี่ยวกับมุกอาหาร”
One-Shot	“คุณคือตลกที่เก่งที่สุดในประเทศไทย นี้คือตัวอย่างมุกตลกที่คุณเคยเล่น	"คุณคือตลกที่เก่งที่สุดในประเทศไทย นี้คือตัวอย่างมุกตลกที่คุณเคยเล่น มุกอาหาร,พริกที่ว่าเผ็ด ยังไม่เด็ดเท่าเรา

	[…A…] จงคิดมุกตลกที่เกี่ยวกับ […X…]”	จงคิดมุกตลกที่เกี่ยวกับอาหาร,"
Few-Shot	“คุณคือตลกที่เก่งที่สุดในประเทศไทย นี้คือตัวอย่างมุกตลกที่คุณเคยเล่น […A…] […B…] […C…] จงคิดมุกตลกที่เกี่ยวกับ […X…]”	"คุณคือตลกที่เก่งที่สุดในประเทศไทย นี้คือตัวอย่างมุกตลกที่คุณเคยเล่น มุกอาหารและการเปรียบเทียบ,พริกที่ว่าเผ็ด ยังไม่เด็ดเท่าเรา มุกตลกของกาแฟ,กาแฟอะไรมีแต่ขี้ฝุ่น หยักใหย่ ก็กาแฟโบราณไง มุกซุป, """ซุปอะไรมีสารอาหารมากที่สุด ซุปเปอร์มาร์เก็ต""" มุกและการกิน,ดอยอะไรกินได้ ดอยคำ มุกภูเขาและความอร่อย,ภูเขาอะไรอร่อย ภูเขาทอง มุกทอดหมู,"""ทอดหมูยังไงไม่ให้ติดกระทะ ทอดในหม้อ""" มุกหมูหัน,จะกินหมูหันต้องเติมอะไร เติมไม้เอก เป็นหมูหั่น ถ้าไม่หั่นก็กินไม่ได้ มุกถั่วงอก,"""ถั่วงอกมีดีอย่างไร ดีกว่ามันไม่งอก""" มุกผลไม้,ผลไม้อะไรจับแล้วเย็น มะอุ่น(องุ่น) มุกผลไม้,ผลไม้อะไรกินแล้วตื่น มะง่วง(มะม่วง) จงคิดมุกตลกที่เกี่ยวกับอาหาร"

For prompting methods above, X is the type of joke and A, B, C are examples of jokes from selected joke type. With Zero-Shot prompting, which is the simplest prompt format, containing only an instruction to create a joke about a specific category (e.g., love, food, animals). For example, the model might be prompted with: “You are the best comedian in Thailand, please come up with jokes about *love*.” (This prompt is translated from the one mentioned in Table II.) Without examples, the model generates a joke solely based on its internal knowledge of humor and the category provided. While efficient, this approach often results in jokes that may lack refinement or relevance to the theme. An example output of Zero-Shot prompting is shown below.

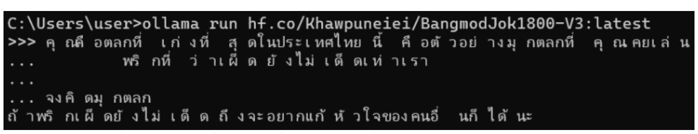

Fig.1. Example output of Zero-Shot prompting

As for One-Shot prompting, we included one example joke. By including one example joke, this method introduces a frame of reference for the model. For instance: “You are the best comedian in Thailand, this is an example from the jokes you created—'Love is not my own but it is yours’--. Please come up with jokes about *love*.” (This prompt is translated from the one mentioned in Table II.) The example helps the model understand the expected humor tone and structure. The generated joke typically exhibits greater thematic alignment and creativity than the Zero Shot method, as the model uses the example as an inspiration.

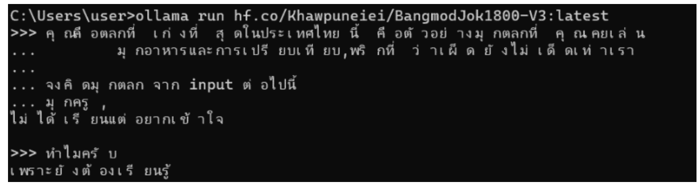

Fig.2. Example output of One-Shot prompting

Finally, Few-Shot prompting enhances contexts by providing multiple example jokes. For example: “You are the best comedian in Thailand, these are examples from the jokes you crated—'You can pick any type of person you want, but if you like my type, come on’, ‘What is the scariest way of flirting? No flirting.’--. Please come up with jokes about *love*.” With multiple examples, the model can identify recurring patterns, themes, and humor mechanics, leading to more sophisticated and contextually appropriate responses.

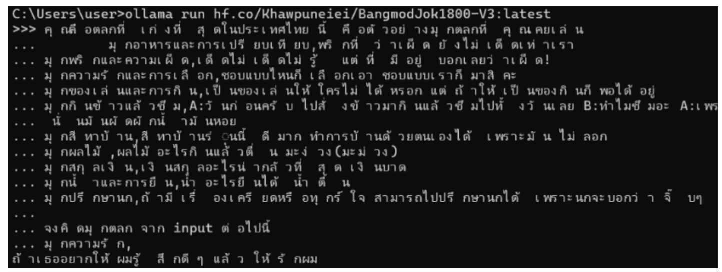

Fig.3 Example output of Few-Shot prompting

B. *Joke Evaluation*

The jokes generated from previous section were subjected to evaluation by a group of 52 undergraduate engineering students from King Mongkut’s University of Technology Thonburi (KMUTT). This demographic, comprising students aged between 18 and 24 years, was selected to ensure spontaneous and natural responses to the humor. No prior training or specific background in humor analysis was required. The demographic characteristics of the participants are as follows:

TABLE III. DEMOGRAPHIC DISTRIBUTION OF PARTICIPANTS

Demographic	*Count (n = 52)*	*Percentage (%)*
Gender		
Male	32	61.54%
Female	20	38.46%
Age (Years)		
Mean	20.01	-
Median	20	-
Standard Deviation	2.52	-
Min - Max	15-35	-

For evaluating jokes, a questionnaire was designed to collect both demographic information and humor evaluation. Each participant evaluated 12 jokes, covering AI-generated jokes (from Zero-Shot, One-Shot, and Few-Shot prompting) and human-created jokes for comparison. The experiment was designed to ensure unbiased evaluation of jokes through three key measures. First, a blind test was implemented, where participants were unaware of whether the jokes were AI-generated or human-created. Second, randomized joke ordering was applied, presenting each joke in a random sequence to eliminate any potential ordering bias. Finally, a binary evaluation method was employed, requiring participants to answer a simple question: "Is the joke funny?" with response options limited to "Funny" or "Not Funny."

TABLE IV. STRUCTURE OF THE QUESTIONNAIRE

Question Type	Response Format	Example Responses
Gender	Multiple Choice	Male, Female
Age	Numeric Entry	18, 19, 20, etc.
Joke Evaluation	Binary (Funny/Not Funny)	Funny / Not Funny
12 Jokes for evaluation (Examples Below)		
- "Why don't provinces like spicy food? Because Phuket ('Phoo-ket' sounds like 'spicy' in Thai)."	Funny / Not Funny	
- "Teachers teach us how to make money, but not how to find love."	Funny / Not Funny	

C. Results and Analysis of Questionnaire data

To ensure the reliability of participant responses in evaluating AI-generated jokes, we computed Cronbach's Alpha (α) [9] as a measure of internal consistency. This statistic assesses whether multiple joke ratings provided by participants exhibit coherence and reliability. The analysis was performed across all 12 joke evaluations collected from 52 participants. The computed Cronbach's Alpha for the dataset was $\alpha = 0.845$, which falls within the 0.775-0.901 range, indicating good internal consistency and providing confidence in the validity of the responses.

To assess the humor effectiveness of AI-generated jokes, participants evaluated each joke using a binary rating system ("Funny" or "Not Funny"). The percentage of jokes rated as funny was calculated for each prompting method and compared against human-generated jokes, which served as the baseline.

TABLE V. PERCENTAGE OF JOKES RATING

Prompting Method	Jokes Rated as Funny (%)	Jokes Rated as Not Funny (%)
Zero-Shot Prompting	16.67%	83.33%
One-Shot Prompting	12.82%	87.18%
Few-Shot Prompting	27.56%	72.44%
Human-Generated (Baseline)	60.90%	39.10%

As seen from Table V, the results indicate that AI-generated jokes performed significantly inferior than human-generated jokes in terms of perceived humor. The Few-Shot Prompting approach yielded the highest percentage of jokes rated as funny (27.56%), outperforming both Zero-Shot and One-Shot Prompting. However, this score remains far below the human-generated joke baseline (60.90%), suggesting that AI-generated humor is still not as effective as human-created humor. In other words, human-generated jokes were rated as funny 60.90% of the time, which is more than twice as effective as Few-Shot AI-generated jokes, reinforcing the gap between AI and human creativity in humor generation.

To determine whether humor perception varied significantly across different prompting methods, a One-Way ANOVA [10] test was conducted. The results indicated a significant difference in humor ratings among the four groups, $F(11,612) = 43.001$, $p = 2.962\times10{-}28$, suggesting that the type of prompting method significantly influenced the perceived humor of AI-generated jokes.

Subsequently, we conducted the post-hoc pairwise comparisons from Tukey's HSD test [11] to interpret the statistical significance which revealed several key differences between the groups. Few-Shot Prompting was found to produce significantly funnier jokes than One-Shot Prompting ($p = 0.0099$), indicating that providing multiple examples enhances humor quality. However, no significant difference was observed between Few-Shot Prompting and Zero-Shot Prompting ($p = 0.0963$), suggesting that a single example does not substantially improve the humor output.

Evidently, human-Generated jokes were rated significantly funnier than all AI-generated jokes, with significant differences observed when compared to Few-Shot ($p = 0.0$), One-Shot ($p = 0.0$), and Zero-Shot ($p = 0.0$). This underscores the dominance of human creativity in generating humor.

Furthermore, no significant difference was found between One-Shot Prompting and Zero-Shot Prompting ($p = 0.8468$), which reinforces the notion that providing a single example does not lead to significant improvements in humor generation compared to offering no examples at all.

TABLE VI. MULTIPLE COMPARISON OF MEANS - TUKEY HSD WITH FAMILY WISE ERROR RATE (FWER) OF 0.05

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	Reject
Few-Shot	Human-Generated	0.3333	0.0	0.2119	0.4547	True
Few-Shot	One-Shot	-0.1474	0.0099	-0.2688	-0.026	True
Few-Shot	Zero-Shot	-0.109	0.0963	-0.2304	0.0124	False
Human-Generated	One-Shot	-0.4808	0.0	-0.6022	-0.3594	True
Human-Generated	Zero-Shot	-0.4423	0.0	-0.5637	-0.3209	True
One-Shot	Zero-Shot	0.0385	0.8468	-0.0829	0.1599	False

As Few-Shot Prompting outputs performed best among AI-generated jokes, we compared their funniness ratings with those of human-generated jokes using an independent t-test [12]. The results revealed a T-statistic of -6.27 and a p-value of 1.195×10^{-9}, indicating a highly significant difference between the two groups. Given that the p-value is much smaller than the threshold of 0.05, we reject the null hypothesis, and this suggests that humor ratings for Few-Shot Prompting are significantly lower than those for human-generated jokes. This implies that, even for the best-performing AI-generated jokes, Few-Shot Prompting still cannot achieve the same level of humor quality as human-generated jokes.

V. DISCUSSION AND CONCLUSION

The results of this study reveal both the progress and limitations of AI-generated humor, particularly when using different prompting methods. While AI has made some advances in humor generation, the findings indicate that it still requires significant improvement to match the quality of human-generated jokes. The analysis of internal consistency, measured using Cronbach's Alpha ($\alpha = 0.845$), shows that participants' responses were highly reliable and consistent across the 12 joke evaluations. This not only supports the validity of the humor ratings but also reinforces the credibility of the evaluation process employed in the study.

One of the key insights from the analysis is the importance of providing AI with sufficient data to improve its performance. As demonstrated in this study, AI-generated humor is highly dependent on the amount of input it receives. Even with the best-performing method, Few-Shot Prompting, AI-generated jokes were rated as funny only 27.56% of the time, which is significantly lower than the 60.90% humor rating for human-generated jokes. This gap suggests that while AI is capable of producing jokes, it has not yet developed the ability to effectively connect with human audiences in the way that human humor does. The results highlight that AI still struggles with the complexity, creativity, and nuance inherent in humor. Human-generated jokes were consistently rated as funnier, indicating that AI has not yet reached a level where it can match, let alone surpass, human creativity in this domain.

Further analyses, including the one-way ANOVA and Tukey's HSD post-hoc test, confirm that the quality of AI-generated humor is strongly influenced by the given data. Few-Shot Prompting, which involves providing the AI with multiple examples, outperformed One-Shot and Zero-Shot Prompting. However, even the best AI-generated jokes still fell significantly short of human-generated humor, suggesting that while more examples can improve AI performance, they are not enough to fully replicate the creativity and cultural depth of human humor. Additionally, no significant difference was found between One-Shot and Zero-Shot Prompting, indicating that providing a single example does not offer a major advantage over no examples at all.

Overall, these findings underscore the critical role that data play in improving AI humor generation. While Few-Shot Prompting shows promises, AI humor still requires further development to approach the quality of human-generated content. As AI's creative capabilities are still in their early stages, future research must focus on providing more diverse, context-rich, and culturally relevant data to enhance its ability to generate humor that resonates more deeply with human audiences. In conclusion, while AI has made progress in generating humor, there are still many challenges to overcome, particularly with regards to cultural and contextual nuances, originality, and the subjective nature of humor. Future research should focus on refining AI models through targeted fine-tuning and integrating human expertise in order to achieve more authentic and effective humor generation.

ACKNOWLEDGMENT

This work was supported by CPE (Computer Engineering Department) Research Acceleration Program, Faculty of Engineering, King Mongkut's University of Technology Thonburi.

REFERENCES

[1] W. A. Terra, The Algorithmic Muse: Generative AI in the Creation of Visual Humour. Master Thesis, 2024.

[2] T. Platt, J. Hofmann, W. Ruch, R. Niewiadomski, and J. Urbain, Experimental Standards in Research on AI and Humor when Considering Psychology, AAAI Fall Symposium - Technical Report, 2012.

[3] D. Gorenz and N. Schwarz, "How funny is ChatGPT? A comparison of human- and A.I.-produced jokes," PLoS ONE 19(7): e0305364, 2024. Available: https://doi.org/10.1371/journal.pone.0305364

[4] A.H. Bower and M. Steyvers, "Perceptions of AI engaging in human expression," Sci. Rep. 2021;11:21181. DOI: 10.1038/s41598-021-00426-z.

[5] H. Avetisyan, P. Safikhani, and D. Broneske, "Laughing out loud – Exploring AI-generated and human-generated humor," 4th International Conference on NLP & Artificial Intelligence Techniques (NLAI 2023). Available: https://ssrn.com/abstract=4707459

[6] J.H. Kim, J.Kim, T.H. Baek and C. Kim, "ChatGPT personalized and humorous recommendations," Annals of Tourism Research, Volume 110, 2025, DOI:10.1016/j.annals.2024.103857.

[7] A. Dubey et al., "The Llama 3 Herd of Models." 2024 [Online]. Available: https://arxiv.org/abs/2407.21783

[8] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models.," in ICLR, 2022 [Online]. Available: http://dblp.uni-trier.de/db/conf/iclr/iclr2022.html#HuSWALWWC22

[9] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, 16, pp. 297–334, 1951.

[10] A. Ross, and V.L. Willson, "One-Way Anova," Basic and Advanced Statistical Tests, Rotterdam, SensePublishers, 2017. Available: https://doi.org/10.1007/978-94-6351-086-8_5

[11] H. J., Keselman and J.C. Rogan, "The Tukey multiple comparison test: 1953–1976," Psychological Bulletin, 84(5), pp. 1050–1056, 1977. Available: https://doi.org/10.1037/0033-2909.84.5.1050

[12] A. Ugoni and B. Walker, THE t TEST: An Introduction, 4, 1995.