wangkongqiang at SemEval-2026 Task 1: MWAHAHA - Competition on Humor Generation

Kongqiang Wang[1], **Peng Zhang**[2], **Qingli Tan**[3]
[1,2]School of Information Science and Engineering, Yunnan University,
[3]School of College of Ecology and Environment, Yunnan University,
Kunming 650500, Yunnan, China
[1]wangkongqiang60@gmail.com, [2]zpp1219@gmail.com,[3]tanqingli@stu.ynu.edu.cn

Abstract

This paper presents our system developed for the SemEval-2026 Task 1: MWAHAHA - Competition on Humor Generation. on Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask A will be conducted in English, Spanish, and Chinese. on Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask B is in English only. To this end, we mainly focus on Subtask A: Text-based Humor Generation in English and Chinese, and Subtask B: Image-Based Caption Generation in English language to use two important languages models: BLIP and Qwen series LLM. For Task B1: Image-only Humor Generation and Task B2: Image and Prompt Humor Generation. Our submission achieved the good ranking place in the test set. All subtasks evaluated using Rating (95% CI) score across different languages and modality contexts. For Subtask A in English and Chinese, Rating score 950 and 1054, 95% CI [922, 982] and [1024, 1104], ranked 16th and 1st respectively. For Subtask B in B1 and B2, Rating score 976 and 987, 95% CI [941, 1007] and [948, 1016], ranked 5th and 3rd respectively. For the final ranking, organizers will use the Rating (95% CI) score. Even so, our approach still has yielded good results.

1 Introduction

While humor understanding has been the focus of many shared tasks, humor generation remains an even more challenging and largely unexplored frontier. *MWAHAHA*, which stands for Models Write Automatic Humor And Humans Annotate, is SemEval 2026's (Ghosh et al., 2026) task 1 (Castro et al., 2026) and is the first task dedicated to advancing the state of the art in computational humor generation. Organizers invite participants to develop systems capable of generating genuinely humorous content under various constraints. For the first time, Semeval organizators introduce a humor generation task, aimed at generate large amounts of genuine and meaningful jokes according to the requirements. The task focuses on the humor generation of multilingual (Muhammad et al., 2025a), multicultural and multimodel content, capturing the complexity of words or news headline across diverse contexts. Participants may participate in one or more of the following two sub-tasks.

Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English, Spanish, and Chinese. Constraints: Each generated joke must respect one of the following constraints, designed to make it difficult to simply retrieve existing jokes from the web. *Word Inclusion*: Must contain two specific words (from a list of rare word combinations). *News Headline*: Must be related to a given news article headline (it could be a punchline, or a joke inspired by the headline).

Subtask B: Multimodal Humor Generation with Images. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. Given an image in GIF format, generate a humorous caption (max 20 words) that enhances its comedic effect, in two variants. *Subtask B1*: Only use the GIF image to inspire the caption. *Subtask B2*: Use the GIF file and complete a given text prompt with humorous content. Based on the humor generation task background of given information text or GIF image, we propose the humor generation method based on the BLIP and Qwen large language model (LLM).

We developed for the SemEval-2026 Task 1: MWAHAHA - A Competition on Humor Generation. on Subtask A: Text-based Humor Generation in English and Chinese. on Subtask B: Multimodal Humor Generation with Images. The code of this

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 80–88

method is available on our GitHub website[1].

2 Related Work

SemEval in previous years has introduced tasks focusing on LLM capabilities and content generation (Vazquez et al., 2025; Ramakrishna et al., 2025; D'souza et al., 2025; Brekhof et al., 2024) to evaluate internal potential elements and potential content of the large language model (LLM). These tasks provided chances with using LLMs and fully leverage their capabilities, which have been extensively utilized for content generation tasks and knowledge question-answering tasks.

2.1 BLIP

The Bootstrapping Language-Image Pre-training (BLIP) model is a unified visual-language pre-training model proposed by Salesforce Research in 2022 (Li et al., 2022). The main objective is to efficiently learn the alignment relationship between images and text, and simultaneously support multiple downstream multimodal tasks.

Let's briefly describe the background of the proposal of BLIP. Before BLIP model, multimodal models usually had two problems: **There was a lot of noise in the text-image data**. The image-caption pairs obtained through web scraping often had mismatches and inaccurate descriptions. **Task fragmentation**. Different tasks (image description, image-text retrieval, VQA) required different structures or different pre-training methods. The core idea of BLIP model: By using the model to continuously generate-filter-relearn high-quality image-text pairs, it achieves bootstrapping multi-modal pre-training.

The following is a detailed introduction to the overall architecture of BLIP model. BLIP consists of three core modules:

- **Image Encoder**. Vision Transformer (ViT) is usually adopted. Convert the input image into a visual token representation.
- **Text Encoder/Decoder**. Based on Transformer. It can be used as a text encoder for understanding tasks. It can also be used as a text decoder for generation tasks.
- **Multimodal Fusion Module**. Achieve the interaction of image and text features through Cross-Attention. Supports multiple task forms.

Key Design: For the same model structure, a unified modeling of understanding and generation is achieved through different attention masks.

Below, we will provide a detailed introduction to the three pre-training objectives of BLIP model.

- **ITC (Image-Text Contrastive Learning)**. Zoom in on the matching image-text. Push away the mismatched text and image pairs. It is used for tasks such as graphic and text retrieval. It is similar to CLIP model, but with cleaner data.
- **ITM (Image-Text Matching)**. Image-text matching classification. Determine whether the image matches the text, this essentially is a binary classification. Introduce hard negative to enhance fine-grained alignment capability.
- **LM (Language Modeling)**. Conditional text generation. Given an image, generate descriptive text. Used for image captioning.

The core innovation of BLIP model: Bootstrapping data cleaning. Bootstrapping mechanism including the following points: Use the initial model to rewrite the noisy text-image pairs. Use the ITM (Image-Text Matching) module to score the generated caption. Select high-confidence image-text pairs. Retrain the model with the high-quality data that has been filtered. Its advantages including: Reduce reliance on manual annotation. Significantly reduce noise interference. Enhance the model's generalization ability.

Below is a summary of the types of tasks that BLIP model can support. BLIP is a universal multimodal foundational model that can be used for: image captioning, visual question answering (VQA), image–text retrieval, visual reasoning, vision–language understanding.

2.2 Qwen Large Language Model (LLM)

In previous studies, the content generated by large language models (LLMs) presents several advantages (Belay et al., 2025). The large language models (LLMs) generate humorous content approach can reduce the errors from manual operation by expanding the learning of the given text content or GIF images. It can make the system generate humorous content more robust. In our study, using

[1] https://github.com/WangKongQiang/SemEval2026-Task-1

the $qwen3-next-80b-a3b-instruct/qwen-flash/qwen-plus$ generative model to perform humorous content generation on the trial dataset and test dataset through prompts to generate large amounts of humorous content that meets the requirements while making use of information from the trial data and the test data during providing prompts for the large language model (LLM). Previous research has demonstrated that large language model (LLM) prompt learning can achieve remarkable success.

In our study, we aim to use multiple qwen series models to assess genuinely humorous content under various constraints. When models are prompted on diverse datasets with different ways, they may produce varied contents on humorous content generation, and detailed revision prompt using different words may improve final generation performance. We use qwen series models mainly from the following models: $qwen3-next-80b-a3b-instruct$, $qwen-flash$, and $qwen-plus$.

3 Methodology

3.1 Overall Architecture

The pursued approach involves using a humorous content generation system composed of the BLIP conditional text generative model and then three different version qwen large language models (LLMs). We used several state-of-the-art natural language processing (NLP) generative models on the field of outstanding large language models (LLMs) in China to create humorous content with different architectures and configurations. We then submission the contents of these models generation using a humorous content generation system to produce the final outputs. We mainly used the following large language models (LLMs) for the outputs: $qwen3-next-80b-a3b-instruct$, $qwen-flash$, and $qwen-plus$.

The following will introduce these generative models one by one. First, let's introduce the $qwen3-next-80b-a3b-instruct$ model. This is a large instruction following model in the qwen3-next series. Its main features are as follows:

- **Architecture and Parameters**: Designed based on the hybrid sparse mixture of experts (MoE), the total parameters are approximately 80 B, but only about 3 B parameters are actually activated per token, thereby achieving extremely high efficiency.
- **Instruction Optimization**: This is a instructor-tuned version that focuses on generating concise responses based on user instructions without creating internal thought trajectories.
- **Ultra-Long Context**: Natively supports ultra-long context such as 256 K tokens, and can reach millions of contexts through expansion.
- **Applicable Scenarios**: Suitable for various tasks such as chatting, complex reasoning, code generation, and reading long documents, especially excelling in long text understanding or large-scale document reasoning.

Summary: An efficient, scalable, and highly generalized instruction following model have large-scale, and low computational cost characteristics.

Second, let's introduce the $qwen-flash$ model. This name is generally used in some API platforms, such as $haijing$ AI/third-party services, representing a practical model variant in the qwen series, whose features include:

- **Hybrid Thinking Mode**: It supports automatic switching between thinking mode and non-thinking mode during conversations, thus being capable of handling both general conversations and a certain degree of complex reasoning.
- **Context Support**: As shown in some API documentation, the 4K context window. There are also parameterized implementations that support longer contexts.
- **Applicability**: Positioned between basic dialogue models and large-scale reasoning models, it is suitable for various text understanding, dialogue and reasoning tasks (Muhammad et al., 2025b).

Summary: A flexible medium-scale model for API usage with emphasizing thought switching and comprehensive performance.

Finally, let's introduce the $qwen-plus$ model. It may have different names on different platforms, but generally refers to the enhanced model of the qwen series that lies between the core large model and the lightweight version. Common features include:

- **Balancing Performance and Efficiency**: Compared with the basic version, it performs

better in reasoning, understanding, and dialogue, while maintaining a balance between speed and cost.

- **High-Long-Term Context**: Some versions support larger contexts such as approximately 131K tokens to handle long text.
- **Applicable Scenarios**: Suitable for scenarios that require comprehensive capabilities such as mathematical reasoning, multilingual interaction, data analysis, code generation, etc. But do not need the computing power of top-tier large models.

Summary: A mid-level general-purpose language model that strikes a balance between capability and cost.

3.2 Implementation Step

For Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English and Chinese for our experiment. We only need two steps to generate the humorous content. First, set the API key. It can get from Alibaba cloud. The API key can be applied from the official website of Alibaba cloud Bailian platform[2].

Second, using the OpenAI client code, enable it to read this tsv file. Based on the columns (id, word1, word2, headline) in the file, automatically invoke the qwen model to generate joke text in Chinese or English and output a new tsv file containing two columns: id and text. Among them, text refers to the generated humorous content.

For Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. We only need four steps to generate the humorous content.

First, we experiment with extract the content of each line from the provided tsv data file, and read the content of the url field to download the corresponding GIF file and save it as a local file,

Second, batch-read all the .gif files in the specified directory from the downloaded local GIF dataset, and uniformly sample each file into a maximum of 5 frames, convert the GIF dynamic pictures to multiple static frames that BLIP model can understand.

Third, use the BLIP model to generate captions for each frame, and automatically merge them into the final caption for the entire GIF dynamic picture. Finally, save the result as a csv file with the following columns: gif_name, final_caption, all_captions.

Fourth, based on the all_captions content of each GIF dynamic picture, call the qwen large language model (LLM) to generate an English humorous title up to 20 words as the final output of Subtask B1: Image-only Humor Generation. Extend the original task B1 to task B2: Image and Prompt Humor Generation. When generating the humorous captions for each GIF dynamic picture, in addition to the all_captions information, the prompt content corresponding to the gif_name is also read from another file, and then added together in the prompt for the qwen large language model (LLM). That is, the model needs to create a joke based on this prompt.

Further, these results will be used submitted to competition on humor generation. After our submit, the system will only compute a basic check to see if our submission is complete. The actual metrics will be calculated later in the pairwise comparison arena.

4 Results and Analysis

4.1 Trial Dataset Analysis

The number of word1, word2 and headline columns in trial dataset are described in Table 1. These texts are mainly used for Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English and Chinese for our experiment. The length and quantity distribution of trial headline text data in English and Chinese are analyzed in Figure 1 and Figure 2 respectively. For Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. This task is further divided into Subtask B1: Only use the GIF image to inspire the caption, and Subtask B2: Use the GIF file and complete a given text prompt with humorous content. The number of url and optional prompt columns in trial dataset are described in Table 2. Distribution of the length of prompt texts for each GIF image in the URL path in Figure 3. It shows the length of prompts relative to each GIF image in the URL path for various cases.

[2]https://bailian.console.aliyun.com/

Table 1: The word1, word2 and headline trial data experiment situation in English and Chinese are described.

Language	Text Column	Count
English	|word1|	100
	|word2|	100
	|headline|	1100
Chinese	|word1|	100
	|word2|	100
	|headline|	900

Language	Text Column	Count
English	|total|	1200
Chinese	|total|	1000

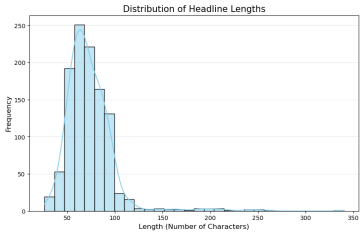

Figure 1: The length and quantity distribution of trial headline text data in English are analyzed.

4.2 Test Dataset Analysis

The number of word1, word2 and headline columns in test dataset are described in Table 3. These texts are mainly used for Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English and Chinese for our experiment. The length and quantity distribution of test headline text data in English and Chinese are analyzed in Figure 4 and Figure 5 respectively. For Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. This task is further divided into Subtask B1: Only use the GIF image to inspire the caption, and Subtask B2: Use the GIF file and complete a given text prompt with humorous content. The number of url and optional prompt columns in test dataset are described in Table 4. Distribution of the length of prompt texts for each GIF image in the URL path in Figure 6. It shows the length of prompts relative to each GIF image in the URL path for various cases.

4.3 Concise Comparative Summary

For the sake of completeness and in an attempt to improve the results obtained by the qwen large language model. For $qwen3 - next - 80b - a3b - instruct$, $qwen - flash$, and $qwen - plus$, the

Table 2: The url and optional prompt trial data experiment situation in Task B1 and Task B2 are described.

Task	Text Column	Count
B1	|url|	1100
B2	|url|	500
	|prompt|	500

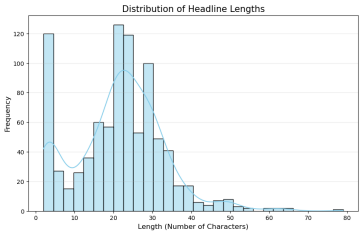

Figure 2: The length and quantity distribution of trial headline text data in Chinese are analyzed.

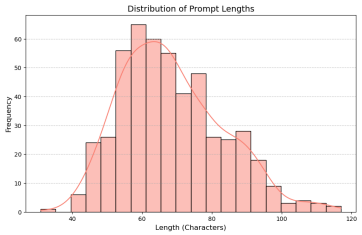

Figure 3: The length and quantity distribution of trial prompt text data in Task B2 are analyzed.

three different model variants were used respectively: See Table 5.

4.4 Trial Dataset Result

The following Table 6 records the official results of SemEval-2026 Task 1: $MWAHAHA$ - Competition on Humor Generation. on Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English and Chinese for our experiment. on Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. Given an image in GIF format, generate a humorous caption (max 20 words) that enhances its comedic effect, in two variants. Subtask B1: Only use the GIF image to inspire the caption. Subtask B2: Use the GIF file and complete a given text prompt with humorous content. The metrics recorded by black bold text is the best (winning) approach in the evaluation task of the trial set for English and Chinese language respectively.

Table 3: The word1, word2 and headline test data experiment situation in English and Chinese are described.

Language	Text Column	Value
English	|word1|	25
	|word2|	25
	|headline|	275
Chinese	|word1|	25
	|word2|	25
	|headline|	275

Language	Text Column	Value
English	|total|	300
Chinese	|total|	300

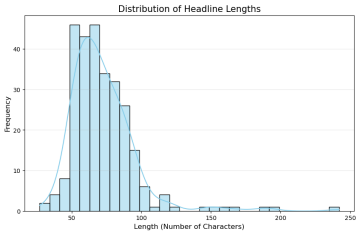

Figure 4: The length and quantity distribution of test headline text data in English are analyzed.

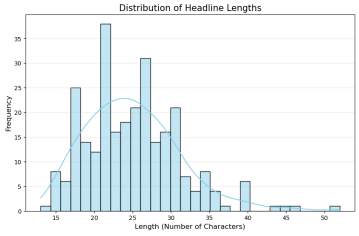

Figure 5: The length and quantity distribution of test headline text data in Chinese are analyzed.

4.5 Test Dataset Result

The following Table 7 records the official results of SemEval-2026 Task 1: $MWAHAHA$ - Competition on Humor Generation. on Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English and Chinese for our experiment. on Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. Given an image in GIF format, generate a humorous caption (max 20 words) that enhances its comedic effect, in two variants. Subtask B1: Only use the GIF image to inspire the caption. Subtask B2: Use the GIF file and complete a given text prompt with humorous content. The metrics recorded by black bold text is the best (winning) approach in the evaluation task of the test set for English and Chinese language respectively.

4.6 Task Prompt Description

From Figure 1 and Figure 2 of the visual analysis, we can observe that 90% of English and Chinese headline in trial dataset, either in the chart or in the

Table 4: The url and optional prompt test data experiment situation in Task B1 and Task B2 are described.

Task	Text Column	Value
B1	|url|	300
B2	|url|	300
	|prompt|	300

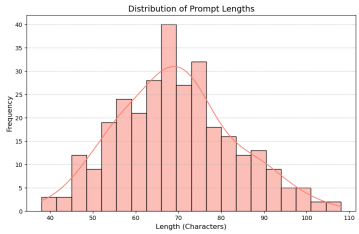

Figure 6: The length and quantity distribution of test prompt text data in Task B2 are analyzed.

previous prompt input column, have no more than 100 and 40 characters respectively. This information could be useful in determining the size of the input tokens for large language models (LLMs), or when the size limit for the tokens consumed by the large language model (LLM) needs to be set.

SemEval-2026 Task 1: $MWAHAHA$ - Competition on Humor Generation. on Subtask A: Text-based Humor Generation. Given a set of text-based constraints, generate a joke. This subtask will be conducted in English and Chinese for our experiment. This task is based on a list of rare word combinations of $word1$, $word2$ or $headlines$ to generative humorous content. For this task, we will create the corresponding prompt using the existing information content, as shown in Table 1. In this case, we should create corresponding prompts in both Chinese and English to ensure they have a certain degree of fault tolerance during the process of generating humor. That is, even when there is missing information, corresponding humorous content can still be generated. The prompt words are shown in the Figure 7 (*Left*). on Subtask B: Image-Based Caption Generation. This subtask explores humor in a multimodal context, combining visual inputs with text generation. This subtask is in English only. Given an image in GIF format, generate a humorous caption (max 20 words) that enhances its comedic effect, in two variants. Subtask B1: Only use the GIF image to inspire the caption. Subtask B2: Use the GIF file and complete a given text prompt with humorous content. The information used to generate humorous content is shown in the Table 2 below. The fundamental difference between these two sub-tasks B1 and B2 lies in whether there is a part of the prompt content for generating humorous content, similar to one being the free creation of humorous content generation, while the other is the continuation of humorous content generation. Overall prompt words for subtasks B1 and B2 are shown in Figure 7 (*Right*).

Table 5: Concise comparative summary for the three different qwen generate large language models.

Model/Variant	Main Positioning	Typical Abilities	Typical Application
qwen3-next-80b-a3b-instruct	|large-scale and efficient MoE instruction model|	extra-long context, high-quality reasoning, and general dialogue	chat assistant, long article analysis, code reasoning
qwen-flash	|medium-sized API model|	flexibly switch between thinking and non-thinking, and maintain stable conversations	daily communication and light reasoning tasks
qwen-plus	|performance and efficiency balance model|	strong comprehensive ability and a longer context	complex dialogues, reasoning and generation

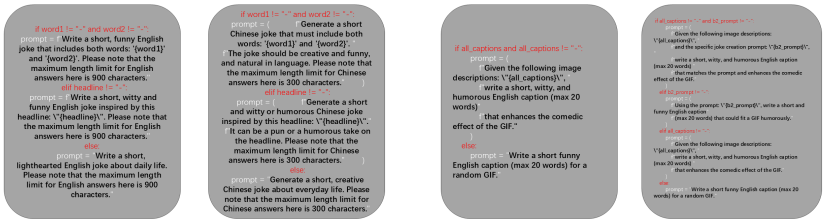

Figure 7: Exhibition of the prompts of texts for each Subtask.

Table 6: The trial dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	Rank	Rating	95% CI	Votes
A	|English|	qwen-plus	2	1004	[966, 1054]	173
A	|English|	**qwen3-next-80b-a3b-instruct**	**2**	**1053**	**[1009, 1103]**	**173**
A	|Chinese|	**qwen-flash**	**1**	**1111**	**[997, 1243]**	**48**
Subtask	**Language**	**Main Technologies**	**Rank**	**Rating**	**95% CI**	**Votes**
B1	|English|	BLIP + qwen-plus	2	1092	[1037, 1176]	1166
B1	|English|	**BLIP + qwen3-next-80b-a3b-instruct**	**2**	**1093**	**[1037, 1177]**	**1175**
B2	|English|	BLIP + qwen-plus	1	1062	[1017, 1336]	1041
B2	|English|	BLIP + qwen3-next-80b-a3b-instruct	1	1062	[1017, 1338]	1040
B2	|English|	**BLIP + qwen-flash**	**1**	**1063**	**[1018, 1337]**	**1040**

Table 7: The test dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	Rank	Rating	95% CI	Votes
A	|English|	**qwen-plus**	**16**	**950**	**[922, 982]**	**-**
A	|English|	qwen3-next-80b-a3b-instruct	-	-	-	-
A	|Chinese|	**qwen-flash**	**1**	**1054**	**[1024, 1104]**	**-**
Subtask	**Language**	**Main Technologies**	**Rank**	**Rating**	**95% CI**	**Votes**
B1	|English|	BLIP + qwen-plus	-	-	-	-
B1	|English|	**BLIP + qwen3-next-80b-a3b-instruct**	**5**	**976**	**[941, 1007]**	**-**
B2	|English|	BLIP + qwen-plus	-	-	-	-
B2	|English|	**BLIP + qwen3-next-80b-a3b-instruct**	**3**	**987**	**[948, 1016]**	**-**
B2	|English|	BLIP + qwen-flash	-	-	-	-

5 Conclusion

Our experiment system employs multiple large language models (LLMs) call from the Bailian platform of Alibaba cloud. These approachs to generate semantic relatedness jokes (Agirre et al., 2014), integrating results from multiple large language model systems: $qwen-plus$, $qwen3-next-80b-a3-instruct$ and $qwen-flash$. The hyperparameter is following: maximum_input_length is 30K, maximum_output_length is 8K, context_length is 32K. The dataset usage is shown in Table 8. Our findings suggest that humor generation semantic relatedness can be deduced from a variety of sources. Although some text generation prompt words (e.g., assume the role of a humor generation expert in the large language model) may not perform as strongly as models specifically designed to obtain humor generation text representations, the results demonstrate that these prompt words, when used in a combined original data manner, can outperform many individual state-of-the-art systems and collaboratively achieve a better correlation with manually inputting to generate humorous content on semantic relatedness (Siino, 2024).

6 Limitation and Future Work

Our experiments are based on English and Chinese language text and multimodel datasets only. Constrained by the size of the language data that the large language model (LLM) has been trained and the availability of online large language models (LLMs), it is regrettable that we did not offer insights into other Asian and African languages (Vaidya et al., 2024). In future research, studies on low-resource languages will be valuable, including tasks such as data collection, annotation, and fine-tune pre-training large language models (LLMs) tailored to these languages.

Acknowledgments

We are very grateful for the assistance and discussions provided by Semeval-2026 Task 1: MWAHAHA - Competition on Humor Generation leaders and organizers.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, page 81–91, Dublin, Ireland. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on*

Table 8: Use dataset supported by Semeval-2026 Task 1: MWAHAHA - Competition on Humor Generation, on Subtask A: Text-based Humor Generation and Subtask B: Image-Based Caption Generation. The style is based on raw data.

Dataset Input	Description	Use or Not
MWAHAHA official dataset	|Datasets for Models Write Automatic Humor And Humans Annotate.|	yes
Other dataset	|Use external or additional corpora.|	no

Computational Linguistics, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Thijs Brekhof, Xuanyi Liu, Joris Ruitenbeek, Niels Top, and Yuwen Zhou. 2024. Groningen team D at SemEval-2024 task 8: Exploring data generation and a combined model for fine-tuning LLMs for multidomain machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Santiago Castro, Luis Chiruzzo, Santiago Góngora, Salar Rahili, Naihao Deng, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Aiala Rosá, Guillermo Moncecchi, J. A. Meaney, Juan José Prada, and Rada Mihalcea. 2026. SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Jennifer D'souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. SemEval-2025 task 5: LLMs4Subjects - LLM-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2570–2583, Vienna, Austria. Association for Computational Linguistics.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025. SemEval-2025 task 4: Unlearning sensitive content from large language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2584–2596, Vienna, Austria. Association for Computational Linguistics.

Marco Siino. 2024. All-mpnet at SemEval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 379–384, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. CLTeam1 at SemEval-2024 task 10: Large language model based ensemble for emotion detection in Hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 365–369, Mexico City, Mexico. Association for Computational Linguistics.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaox-

iong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.