

Howard University-AI4PC at SemEval-2026 Task 1: Exploring Prompt Strategies for Automatic Humor Generation

Abdulmujeeb Lawal and Saurav K. Aryal*

Howard University

AI4PC Lab

lawalmujeeb78@gmail.com

sauravv.aryal@howard.edu

Abstract

We present our solution system for SemEval-2026 Task 1-Subtask A, a humor generation task requiring systems to generate jokes, given either a news headline or word-pair inputs. Our approach used the Llama-3.1-8B-Instruct model which we selected after comparing several candidate models and humor strategies across our experiments. For the headline inputs, we used a two-shot prompt to frame the output as a *tweet* and specifying the tone proved to be a particularly important factor in output quality. As for the word-pair inputs, we instructed the model to commit to an everyday situation and generate a *funny thought* based on that. Also, while experimenting, we noticed that models would start a joke one way with the first word and abruptly shift context mid-joke just to include the second word, and committing to a single situation helped handle that. We also made use of personas here, specifically using Dave Chappelle, following prior work on persona-based prompting for multilingual generation (Ince and Aryal, 2025; Aryal and Pant, 2025,?). Our final system shared 2nd place with 3 other systems out of 32 total systems and achieved an Elo score of 1020. Achieving these results, with no fine-tuning, suggests that careful prompt design alone can yield competitive results.

1 Introduction

Humor is a very unique human phenomenon in the sense that it appears in almost all aspects of our social life. Beyond its large role in entertainment, research has shown that humor also plays an important role in educational settings. When used appropriately, it increases engagement and can make complex material more approachable (Zhou and Lee, 2025). As AI systems and chatbots become more embedded in how we communicate and learn,

the ability to generate humor becomes more meaningful and important.

However, building humor generation systems is not straightforward. Humor is cognitively complex since jokes often require the listener to draw on some background knowledge they may or may not have. To compound this, humor is also deeply subjective as what is considered funny, appropriate or offensive differs across cultures and individuals (Jiang et al., 2019; Ngueajio et al., 2025; Aryal et al., 2023a,a). These properties make humor generation one of the more challenging problems in NLP.

SemEval-2026 Task 1 (Castro et al., 2026; Aryal and Pant, 2025; Rijal and Aryal, 2025; Ince and Aryal, 2025) directly addresses this problem by providing a shared benchmark for humor generation, inviting systems to tackle some of these difficulties in a comparable setting. In this paper, we describe our participation in Subtask A of the task and investigate how prompt design and structured instructions, can help guide LLMs toward more effective humor generation. Our final system was tied for 2nd on the official Subtask A English test set.

2 Background and Related Work

Some of the earliest work in computational humor generation can be traced back to the 90's. Back then, humor systems were mostly built around fixed joke templates filled using lexical databases like WordNet and ConceptNet to create riddles and puns (Amin and Burghardt, 2020). Over time, the field shifted towards neural approaches with seq2seq models and more recently LLM's which allowed for more creative joke generation. Recent SemEval systems further demonstrate the effectiveness of prompt-only and lightweight adaptation strategies across multilingual NLP tasks (Aryal and Pant, 2025; Rijal and Aryal, 2025; Aryal and Pant, 2025;

*Corresponding author

Ince and Aryal, 2025). However, despite all these advances, modern LLMs still struggle to match human-level humor generation. (Loakman et al., 2025; Aryal and Pant, 2025; Ngueajio et al., 2025).

(Amin and Burghardt, 2020; Aryal and Pant, 2025; Aryal and Adhikari, 2023) went further to identify incongruity - lack of harmony - as the central mechanism which underlies most humor generation approaches and it’s a principle that directly informed some of our earlier experiments in the task. In 2025, (Loakman et al., 2025) demonstrated that LLMs struggle especially with jokes requiring cultural knowledge — precisely the type of challenge the news headline constraint of our task poses.

3 Task and Data Description

SemEval-2026 Task 1, *MWAHAHA (Models Write Automatic Humor And Humans Annotate)* (Castro et al., 2026) contained two subtasks. We limited our participation to the first one, Subtask A, which required systems to generate a joke based on one of two text constraints: a news headline, where the joke must be inspired by or react to the given headline, and a word-pair, where the joke must include the two given words which were selected from a list of rare word combinations.

This constraint was intentional as tying each joke to a specific input was expected to make it difficult for systems to simply retrieve popular jokes from the internet. The subtask was also multilingual, covering languages like English, Spanish, and Chinese; though we restrict our participation to English.

Type	Input	System Output
Headline	What the experts want you to know about these key ADHD symptoms	Experts want to educate you on ADHD symptoms, because what we really needed was another thing to Google at 3am.
Word-Pair	<i>drill, book</i>	You ever notice how drill teams practice their routine in the library? Right next to the ‘Don’t make a sound’ book.

Table 1: Actual inputs and our system outputs in Subtask A.

The final-evaluation english dataset contained 300 inputs. These were largely skewed toward headlines, with 275 headline inputs and only 25

word pairs. Also, the organizers provided no labeled data and participants were given only inputs, with no reference jokes or annotations.

Evaluation for this task was done using pairwise rankings. Annotators were members of the general public, and were presented with two outputs for the same input. They were to then select the funnier one, or record a tie when the outputs were nearly identical. Participants were ranked using an Elo-based leaderboard, where systems sharing a rank had overlapping 95% confidence intervals, meaning they were statistically indistinguishable. For example, three systems could share rank 1, despite having different Elo scores, because their confidence intervals overlap. This meant that Elo scores and final ranks did not always directly correspond.

Finally, the organizers provided an official baseline - which also participated in the competition - using Gemini 2.5 Flash based on simple prompts¹ for each input type.

4 System Description

We built our system using Llama-3.1-8B-Instruct², consistent with prior prompt-based LLM systems in SemEval settings (Aryal and Pant, 2025,?; Ince and Aryal, 2025) as it was more consistent in testing than other models we tried: Qwen-2.5 and Gemma 7B. We did not fine-tune the model and relied entirely on prompting throughout. All experiments were run on an A100 GPU on Google Colab and, as for model parameters, we used a temperature of 0.7 and 80 token output limit for both constraint types. We also limited the number of generated characters to 900 to avoid unnecessarily long jokes.

Setting	Headline	Word Pair
Model	Llama-3.1-8B-Instruct	
Temperature	0.7	
Max Tokens	80	
Max Characters	900	
Generations	1	
Prompt Type	Few-shot	Zero-shot
Shots	2	0
Retry Limit	1	3

Table 2: System configuration per input type. Top rows are shared; bottom rows differ.

¹https://pln-fing-udel.ar.github.io/semEval-2026-humor-gen/baseline/baseline_prompts.txt

²<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

4.1 Headline Inputs

For the headline inputs, we used a two-shot [prompt](#) asking the model to write a *tweet* reacting to the headline. More specifically, we had the [prompt](#) ask for an observational type *tweet* and a slightly cynical tone. We chose this method because during development we observed that *tweet-style* jokes seemed to be more relatable than other formats like one-liners and Question-Answer jokes.

4.2 Word-pair inputs

For the word-pair inputs, instead of asking the model to create a joke, we asked the model to use everyday observations to create a *funny thought* instead - something a comedian would observe, say out loud, and an audience could relate to. We also had the model assume the persona of Dave Chappelle, after trying a number of other comedian personas.

The challenge with this method, however, was the fact that the model would at times struggle to include both words in the joke but when it did they were sometimes decently funny. We found no good way to handle this without reducing the joke quality and decided to implement a retry mechanism to regenerate the joke if both words were not included.

4.3 Generation Protocol

For both input types, we generated a single output for each input without any post-ranking or filtering. For the headline, we truncated any outputs that were above 900 characters. For the word pairs, we allowed up to 3 generation attempts, stopping as soon as both words appeared in the input. We also kept the last input in case, for some reason, all words were not included. This retry mechanism definitely worked all we were able to verify that all the 25 word-pair outputs in our results contained both words.

Finally, we also set no random seed, meaning our results were non-deterministic.

5 Experimental Setup

While experimenting, we came up with a simple scoring system where we assigned a point to any joke which we felt could elicit some kind of reaction - including a chuckle. This led us to drop Gemma early on as it kept producing very bad jokes and eventually drop Qwen soon after since the jokes just could not keep up with those Llama generated.

An issue we had very early on was that virtually all the models defaulted to Question-Answer type jokes which led to outputs like: “*Why did the X do Y? Because Z*”. These jokes were not necessarily funny and suppressing this behavior required our prompts to be increasingly explicit while testing. For the headline inputs, we tried a number of formats including simple one liners, stand up monologue and even Dark Humor, before settling on the *tweet* format. Here, we observed that including examples in the prompt did help but what seemed to create better results was asking for a specific tone.

As for the word-pairs, besides from settling on a good persona, the main challenge was making the jokes coherent. Models would start the joke one way, with the first word, and then shift direction completely just to include the second word and this made for absurd sounding jokes. So we told the model to commit to a single *situation* instead. Finally, for both inputs, we played with temperature, trying out 0.5, 0.7 and 0.9. Using 0.5 led to more *statements* than actual jokes and 0.9 led to a lot of rambling, which were both expected.

To support reproducibility, our implementation is publicly available on GitHub³. All experiments were run on an NVIDIA A100 GPU via Google Colab using Llama-3.1-8B-Instruct⁴. Also, the complete prompts used in our experimental setup are provided in the [Appendix](#). No random seed was set during generation, so exact outputs may vary across runs.

5.1 Limitations and Safety

Our system has some limitations worth acknowledging. The use of a comedian’s persona and an observational tone carries the risk of creating outputs that may be perceived as offensive in some cases. We applied no explicit toxicity filtering, so jokes were as is. Additionally, the system was designed solely for research purposes and was not deployed in any real-world setting.

6 Results

Systems were evaluated using Elo-based ranking where scores went up or down depending on how many pairwise comparisons a system could win

³<https://github.com/lawal-mj/semEval-2026-task-1>

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

Rank	System	Elo	95% CI
1	baseline	1081	[1045, 1110]
1	SLPG_FJWU_Insa	1080	[1046, 1120]
1	berkbubus	1079	[1057, 1115]
1	jct_sb	1063	[1036, 1099]
1	jjuliar	1060	[1027, 1091]
1	mrshu	1045	[1018, 1073]
1	lmfaoooo	1041	[1009, 1064]
1	begumyivli	1041	[1008, 1068]
1	oseminck	1034	[1005, 1072]
2	t_abbasi7	1029	[1001, 1053]
2	sinaeskandari	1022	[989, 1054]
2	abdulmj1 (AI4PC)	1020	[992, 1053]
3	arampageos	1019	[984, 1045]
2	stefanoprea	1012	[986, 1051]
4	zhangxulong	1012	[985, 1036]
...			
31	hemeshkumar_31	843	[802, 875]

Table 3: Leaderboard results for SemEval-2026 Task 1, Subtask A (English). Our system is shown in bold.

against other systems. Each rating also got a *Confidence Interval (CI)* and when two systems shared a rank, that meant their intervals overlapped and one could not confidently say which system was significantly better.

Our System **abdulmj1** - may be renamed as **AI4PC** - shared a rank of 2nd with 3 other systems out of 32 total systems, with a score of **1020** and a 95% CI of **[992, 1053]** on the final leaderboard. Nine other systems shared rank 1, all with higher confidence intervals than ours, indicating they’d be better.

The official baseline, based on Gemini 2.5 Flash, achieved an Elo score of **1081** and ranked higher than all the systems in the sub task. While our system does fall below baseline, the gap is relatively modest with the lowest rank 1 system having an Elo score of **1034** as opposed to our **1020**. This still suggests that our system was relatively competitive.

Lastly, there were no per-constraint results provided so we are unable to report whether we did better at the word-pair inputs or the headline inputs.

7 Analysis

While analyzing our results, we observed that the quality of headline inputs varied significantly. Stronger jokes were more concise and landed through irony for the most part, or were something everyone could relate to, like the *LeBron* joke below referencing student loans. Weaker jokes failed to understand the headline and usually created a more generic response like the *blender* example provided below.

The word-pair jokes were even more inconsis-

Quality	Input	Output
Strong	LeBron James to return for 23rd NBA season	LeBron James still has a longer career than my student loans.
Weak	Ninja dual zone air fryer is at its lowest price for Black Friday 2025	Because what we really needed, a way to cook food while also perfecting our stealth skills.

Table 4: Actual system output examples of strong and weak humor generation from news *headlines*.

Quality	Words	Output
Strong	measure, fridge	You know what’s wild? When you finally measure the leftovers in the fridge and realize you’re 3 days too late.
Weak	roll, shirt	You ever notice how when you’re at the laundromat, the wash cycle is so long, you start to roll up your shirt, hoping to get in a few extra minutes of folding time.

Table 5: Actual system output examples of strong and weak humor generation from *word-pairs*.

tent. A recurring issue was that the model latched onto the prefix “You know what...” suggesting a learned template pattern rather than varied observations. Additionally, the retry mechanism we employed ensured both words were always present in the final output (see Section 4.3), though word inclusion alone did not guarantee humor quality as seen in the *roll/shirt* example.

Overall, given the fact that 275 of the 300 were headlines, we suspect that our overall Elo rating was most likely driven by headline performance, as the word-pair inputs would have had limited impact on final ranking.

8 Conclusion

We approached the challenge using Llama-3.1-8B-Instruct⁵, and for both input types we employed no fine tuning. For the headline inputs, we had the model create jokes in the form of a *tweet*, while for word-pairs, we had the model act as a comedian, Dave Chappelle, commit to an everyday situation and try to fit in both words in a funny thought.

Consequently, our system shared a rank of 2nd with 3 other systems out of the total 32 systems

⁵<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

involved and had an Elo score of 1020. Achieving this without any fine-tuning suggests that prompt design alone can be a competitive strategy for humor generation (Aryal and Pant, 2025; Ince and Aryal, 2025; Aryal et al., 2023b).

Word-pairs remained our major challenge as the models tended to shift the situation of the joke unnecessarily just to include the second word as required.

For future work, we plan to look into some fine-tuning to see if that yields better performance. We also explored Graph RAG a bit, with the aim of giving the models more context about the word usage, but the systems tended to lean heavily into the word definitions and its relationships rather than the actual creative humor. We also hope to look further into employing an *LLM-as-a-Judge* system. We did try it briefly, but found no objective way to have the LLM identify the funnier joke most of the time. Lastly, we plan to evaluate the approach on larger models to assess whether scale improves humor quality.

References

- Miriam Amin and Manuel Burghardt. 2020. *A survey on approaches to computational humor generation*. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.
- Saurav Aryal and Kritika Pant. 2025. Howard university-ai4pc at semeval-2025 task 9: Using open-weight bart-mnli for zero shot classification of food recall documents. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1919–1923.
- Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023a. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.
- Saurav K Aryal, Howard Prioleau, Surakshya Aryal, and Gloria Washington. 2023b. Baseline performance for multilingual codeswitching sentiment classification. *Journal of Computing Sciences in Colleges*, 39(3):337–346.
- Saurav Keshari Aryal and Gaurav Adhikari. 2023. Evaluating impact of emoticons and pre-processing on sentiment classification of translated african tweets.
- 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: MWA-HAHA, Models Write Automatic Humor And Humans Annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.
- Amir Ince and Saurav Aryal. 2025. Howard university-ai4pc at semeval-2025 task 11: Combining expert personas via prompting for enhanced multilingual emotion analysis. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1645–1655.
- T Jiang, H Li, and Y Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10:123.
- Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who’s laughing now? an overview of computational humour generation and explanation. *Preprint*, arXiv:2509.21175.
- Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys*, 57(7):1–37.
- Suprabhat Rijal and Saurav Aryal. 2025. Howard university-ai4pc at semeval-2025 task 7: Crosslingual fact-checked claim retrieval-combining zero-shot claim extraction and knn-based classification for multilingual claim matching. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1777–1782.
- W Zhou and J C Lee. 2025. Teaching and learning with instructional humor: a review of five-decades research and further direction. *Frontiers in Psychology*, 16:1445362.

Appendix

Headline Prompt

Task: Write a single joke in the form of a tweet reacting to the headline below.

If the headline is a question, answer it sarcastically or ironically.

Tone: Observational, clever, slightly cynical.

Constraint:

RETURN ONLY THE TWEET

No quotes

No emojis

Max 280 characters

Example 1:

Headline: Scientists say we may live on Mars one day

Tweet: We can't get people to return shopping carts, but yeah, let's trust them with a whole new planet.

Example 2:

Headline: A study finds that working longer hours increases stress

Tweet: Wow. Next you're gonna tell me fire is hot and my job doesn't love me back.

Headline: {headline}

Tweet:

Word Pair Prompt

You are writing observational stand-up comedy in the style of Dave Chappelle:

sharp, grounded, and based on everyday human behavior.

Task:

Write ONE short, self-contained funny thought that naturally includes BOTH "{word1}" and "{word2}".

Before writing the joke, silently do the following:

1. Pick ONE clear, everyday situation where BOTH words logically belong.
2. Commit to that situation completely.
3. Do NOT change scenarios, settings, or topics mid-joke.

Rules:

- Under 20 words total
- MUST include "{word1}" and "{word2}"
- ONE situation only (no pivots, no escalation to unrelated ideas)
- Both words must feel inevitable, not forced
- Observational, relatable, and punchy
- The humor should come from recognition, not randomness

Disallowed:

- Random word insertion
- Abstract philosophy
- Sudden topic shifts to accommodate the second word
- Multiple mini-jokes

Write only the final joke.

Funny thought: