Text is not all you need:

Multimodal Prompting Helps LLMs Understand Humor

Ashwin Baluja, Northwestern University

The task: humor explanation

Given an input joke, output a natural language explanation of the joke:

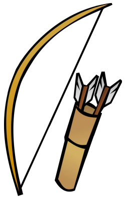
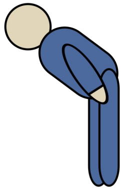
Joke: `My neighbor's sprinkler is a constant irrigation to me!`

Explanation: `This is a pun on 'irritation' which is the state of feeling annoyed, impatient, or slightly angry...`

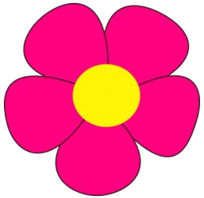
Humor explanation, focusing on puns

Why puns?

- Puns exploit ambiguities in input modality
  A. Homographs: same spelling, different meanings (~ambiguous audio)
  B. Heterographs: different spellings, similar sounds (~ambiguous text)

A. `bow vs. bow`




B. `flower vs. flour`

Existing approaches and related work

1. LLM-based joke explanations
   – Xu et al. "A good pun is its own reword": Can Large Language Models Understand Puns?."
2. Fine-tuned LLMs for recognizing types of humor
   – Wu et al. "Humour classification by fine-tuning LLMs: CYUT at CLEF 2024 JOKER Lab subtask humour classification according to genre and technique."
3. Fused modality representations
   – Hasan et al. "Humor knowledge enriched transformer for understanding multimodal humor."
4. Paired modality training
   – Liu et al. "Visual instruction tuning."

Agenda

1.

How to provide the LLM with information to preserve the ambiguity in puns?

2.

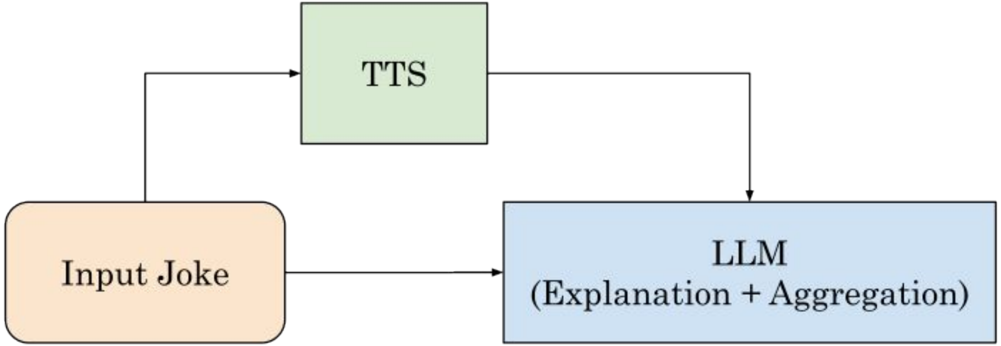
How does this method perform across different datasets, types of puns, and humor in general?

3.

What is the underlying mechanism behind the performance increase?

Multimodal prompting strategy

- LLM takes in both audio and text simultaneously
- `OpenAI tts-1-hd` used for TTS
- `Gemini-1.5-Flash` used for generating explanations



Prompt composition

- Task in prompt is pun recognition
- Chain-of-thought reasoning is used as explanation
- Specific wording needed match output style to dataset

Definition of a pun
Instructions
    detect whether pun or non-pun,
    describe why,
    don't address modality
Examples
    joke text, joke detection, reasoning
    … x 6
==================================
Input joke
Input audio

Datasets tested (puns)

- **SemEval-2017 task 7**
  - annotations and human joke explanations
  - Miller et al. “SemEval-2017 task 7: Detection and interpretation of English puns.”
- **Context Situated Puns**
  - annotations, no human joke explanations
  - Sun et al. “Context-Situated Pun Generation

annotation

explanation

```
"het_1105": {
    "pun_word": "barbarously",
    "pun_sense": "in a barbarous manner",
    "alter_word": "barber",
    "alter_sense": "a hairdresser who cuts hair and shaves
        beards as a trade",
    "human_text": "' ' Give me a haircut , ' ' Tom said
        barbarously .",
    "human_explanation": "The joke is a play on words. To
        do or say something 'barbarously' is to be loud or
        rowdy.  'Barbarously' sounds like 'barber', and
        barbers cut hair."
}
```

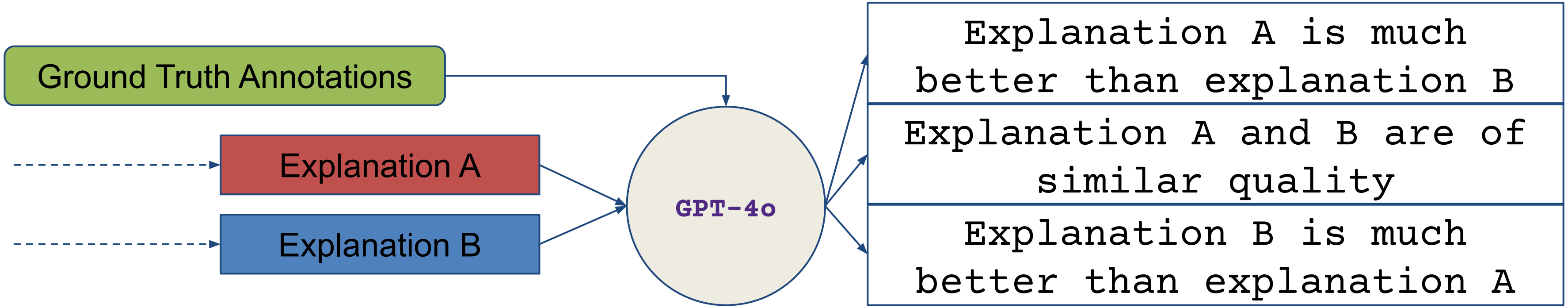
Datasets tested (other)

- **ExplainTheJoke**
  - Many types of jokes
  - Inconsistently formatted joke explanations
  - Explanations summarized by LLM into consistent format, used as annotation
  - https://explainthejoke.com/

{
"joke": "Q: What do you call a grilled cheese sandwich that gets right up in your face? A: Too close for comfort food!,",
"explanation": "This joke is funny because it plays on the double meaning of 'too close for comfort,' using the idiom to refer to both physical proximity and emotional closeness, while also referencing the comforting nature of food."
}

Evaluation methods

- Comparing natural language outputs
- **`GPT-4o`** used as a judge
- Annotations provided to judge to ground decisions
  - definition and spelling of both interpretations of the pun
- Judged by pairwise comparison win-rates



Results (puns)

SemEval

	Heterograph		Homograph	
	Win %	Tie %	Win %	Tie %
Baseline	47.76	5.64	68.89	8.40
with audio	**51.74**	4.56	**72.59**	6.36

vs. human explanations

Results (puns)

Context Situated Puns

	Heterograph		Homograph	
	Win %	Tie %	Win %	Tie %
Baseline	33.87	29.65	35.08	28.08
with audio	**36.49**		**36.85**	

vs. each other

Results (other)

ExplainTheJoke

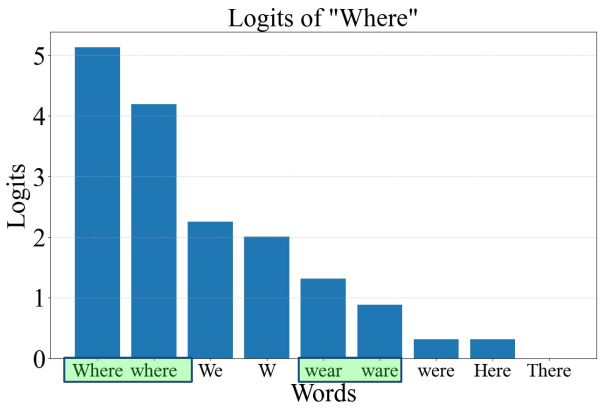
	Win %	Tie %
Baseline	12.81	71.75
with audio	**15.44**	

vs. each other

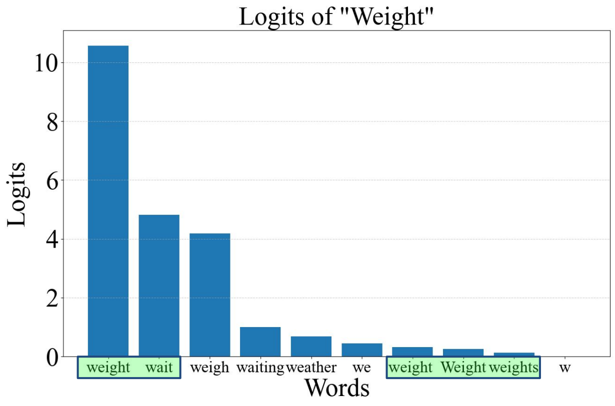
Insights: audio logits

In each example, an audio-capable LLM was asked what word was spoken at a position in an audio clip.

“where”



“patience is a heavy weight”



Insights: TTS parameters

	Heterograph		Homograph	
	Win %	Tie %	Win %	Tie %
Nova (female)	44.59	5.25	71.48	6.30
Onyx (male)	45.44	5.65	73.33	6.42
Alloy (androgynous)	**51.74**	4.56	72.59	6.36
Onyx + Alloy	47.91	3.79	**73.09**	5.74

- Multiple voice types tested, including passing in more than one voice at a time
- No clear performance trend with voice type

Extensions and future work

1. Expand additional input modalities beyond audio
2. Benchmark new tasks, both outside of puns and humor
3. In-depth analysis of effect of TTS parameters on performance across types of jokes, subject matter

Q: `What's in the middle of the pacific?`

A: `"c"`

Conclusions

1.

Multimodal prompts improve humor understanding

2.

Significant performance improvements are possible with no additional training

3.

This method paves the way for broader applications, both with audio input or with other modality inputs

References

Sajal Aggarwal et al. 2023. Multimodal sarcasm recognition by fusing textual, visual and acoustic content via multi-headed attention for video dataset. In 2023 World Conference on Communication & Computing (WCONF), pages 1–5.

Salvatore Attardo and Lucy Pickering. 2011. Timing in the performance of jokes. HUMOR, 24(2):233–250.

Tom Brown et al. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Chiara Bucaria. 2004. Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines. HUMOR, 17(3):279–309.

Wei-Lin Chiang et al. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. Preprint, arXiv:2403.04132.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint, arXiv:2403.05530.

Md Kamrul Hasan et al. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 12972–12980.

Dan Hendrycks et al. 2021a. Aligning AI with shared human values. In Proceedings of the International Conference on Learning Representations (ICLR).

Dan Hendrycks et al. 2021b. Measuring massive multitask language understanding. In Proceedings of the International Conference on Learning Representations (ICLR).

Chris Hua. 2024. Gazelle v0.2.

Haotian Liu et al. 2023. Visual instruction tuning. In NeurIPS.

References (cont.)

Tristan Miller et al. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o system card. Preprint, arXiv:2410.21276.

Minghao Shao et al. 2024. Survey of different large language model architectures: Trends, benchmarks, and challenges. IEEE Access, pages 1–1.

Jiao Sun et al. 2022. Context-situated pun generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4635–4648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Theblackcat102. Theblackcat102/joke_explaination - datasets at hugging face.

Peiyi Wang et al. 2024. Large language models are not fair evaluators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Caleb Warren et al. 2021. What makes things funny? An integrative review of the antecedents of laughter and amusement. Personality and Social Psychology Review, 25(1):41–65. PMID: 33342368.

Colin White et al. 2024. LiveBench: A challenging, contamination-free LLM benchmark. Preprint, arXiv:2406.19314.

Shih-Hung Wu et al. 2024. Humour classification by fine-tuning LLMs: CYUT at CLEF 2024 Joker Lab subtask humour classification according to genre and technique. In Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, pages 1933–1947.

Zhijun Xu et al. 2024. "A good pun is its own reword": Can large language models understand puns? Preprint, arXiv:2404.13599.

Rowan Zellers et al. 2019. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Lianmin Zheng et al. 2024. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23, Red Hook, NY, USA. Curran Associates Inc.

Thank you.

Ashwin Baluja, Northwestern University

baluja@u.northwestern.edu