Automated Audio Captioning with Knowledge Infusion

Project 1.1 - Group 16



Maastricht University

Tom Smit, Adrien Bersia, Enders Turkers, Dumitru Versebeniuc

Automated Audio Captioning





AAC Model

Caption

AAC:

“Dog” → “A dog is barking angrily”

Hallucinations





AAC Model

Caption



“A dog is flying through the air”

Ontology

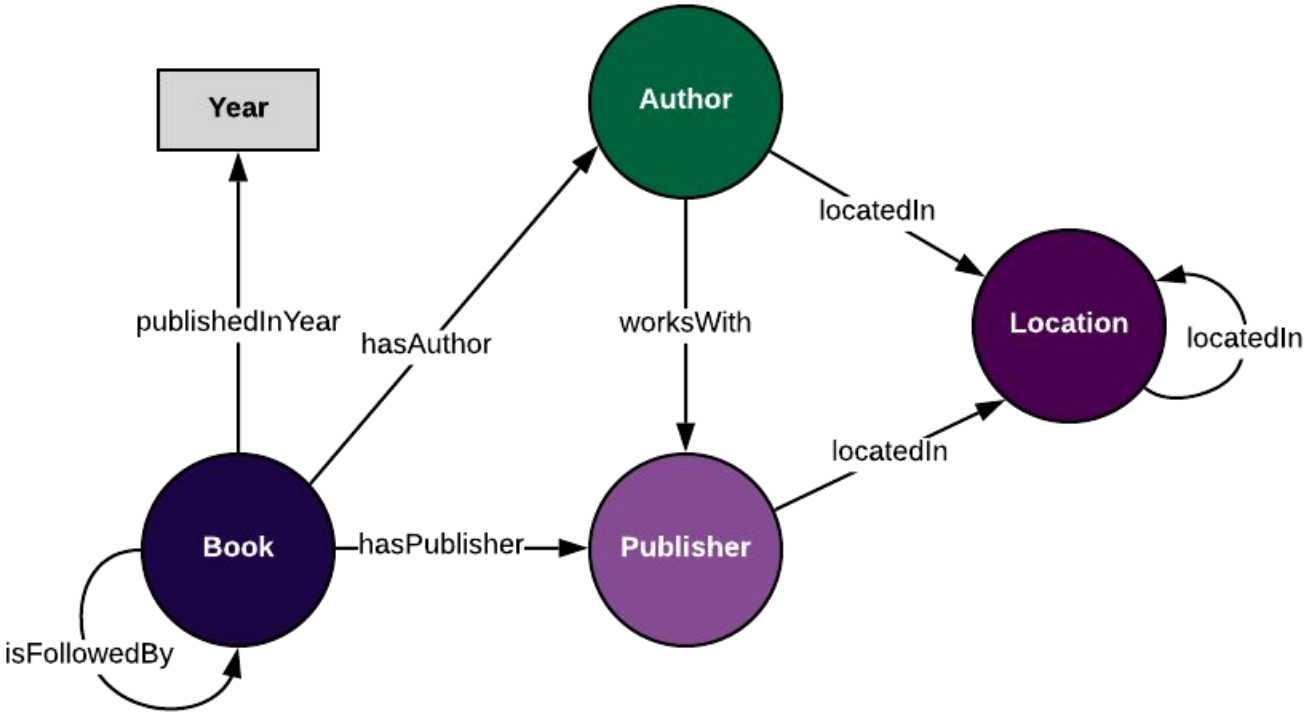


Image from: https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/

General Idea



AAC Model

Caption

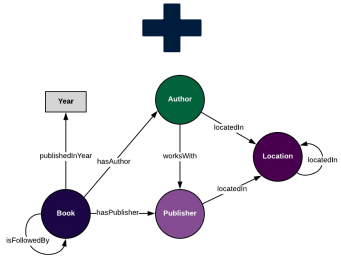




AAC Model

Caption





Problem Statement

- **Can a neuro-symbolic model help in assisting the audio captioning model to generate a more factual caption?**

Research Questions

Research Questions

1. What type of knowledge infusion contributes the most in reducing hallucinations and improving the accuracy and relevance of generated captions in our model?
2. What multimodal method allows for an optimal fusion of our ontology model and audio encoder?
3. What are the most robust and insightful evaluation metrics for assessing the quality, relevance, and factual correctness of the captions generated by the proposed model?

Outline

- Datasets
- Baseline models
- Knowledge infusion
- Ontology
- Ontology based methods
- Evaluation
- Hallucination detection
- Discussion and Conclusions

Datasets

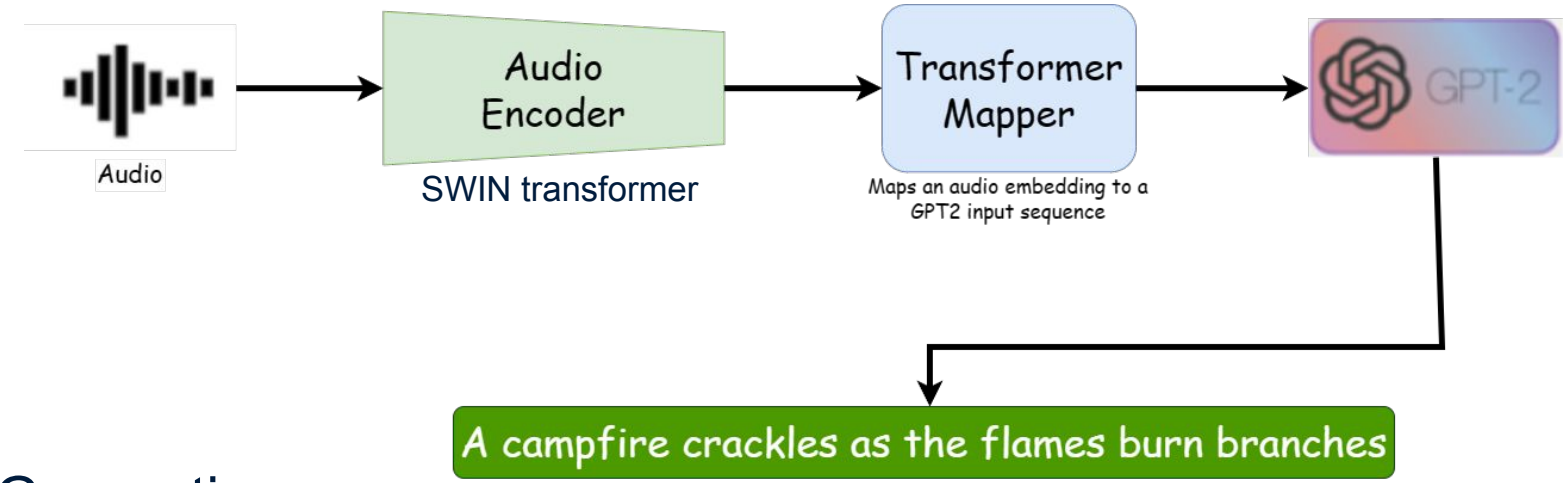
Dataset

- **Clotho v2.1 - 7K audios**

- Each audio sample has five human-labeled captions (a total of 34 870 captions).
- Audio samples are of 15s to 30s duration and captions are 8 to 20 words long.
- Topic balanced splits.

Microsoft’s CLAP 👏

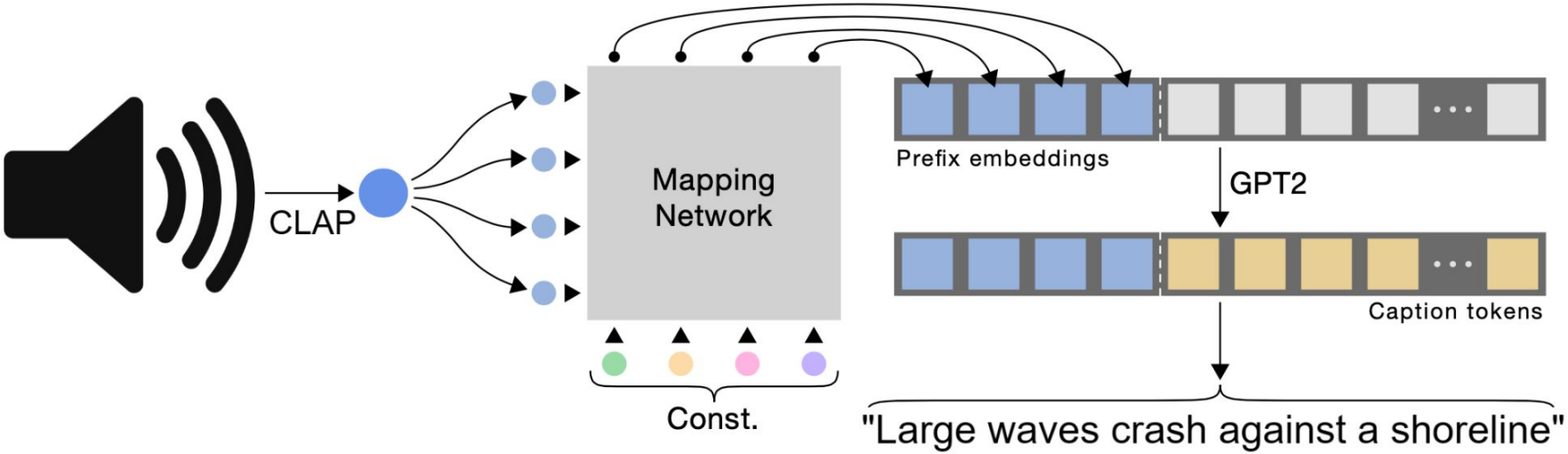
Microsoft CLAP👏



Caption Generation:

- Waves crash against a shoreline as the wind blows.
- A tractor is moving slowly and then slowing down.

MSCLAP's Transformer Mapper

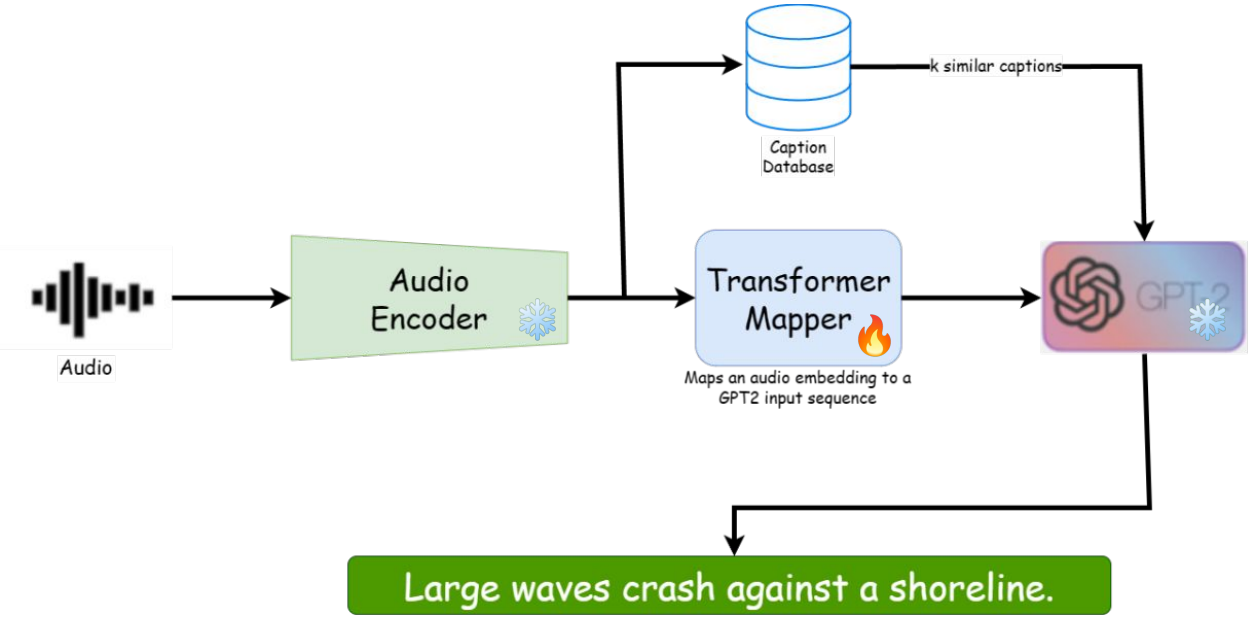


- Increasing the number of these vectors (the prefix size) enhances the system's effectiveness.
- CLAP team discovered that the best results are achieved when the LM remains static and the transformation network is a transformer with 8 layers of multi-head self-attention, each containing 8 heads.

Knowledge Infusion

Retrieval Augmented Generation (RAG)

LangChain



Prompt Example:

Similar audios sound like

Waves constantly crash loudly and with force against shore.
Heavy waves crashing, with a single, quick clang at the end.
Waves are crashing loudly while water is splashed.
Heavy waves are crashing with a single quick clang at the end.
Waves crashing loudly into the shore multiple times.

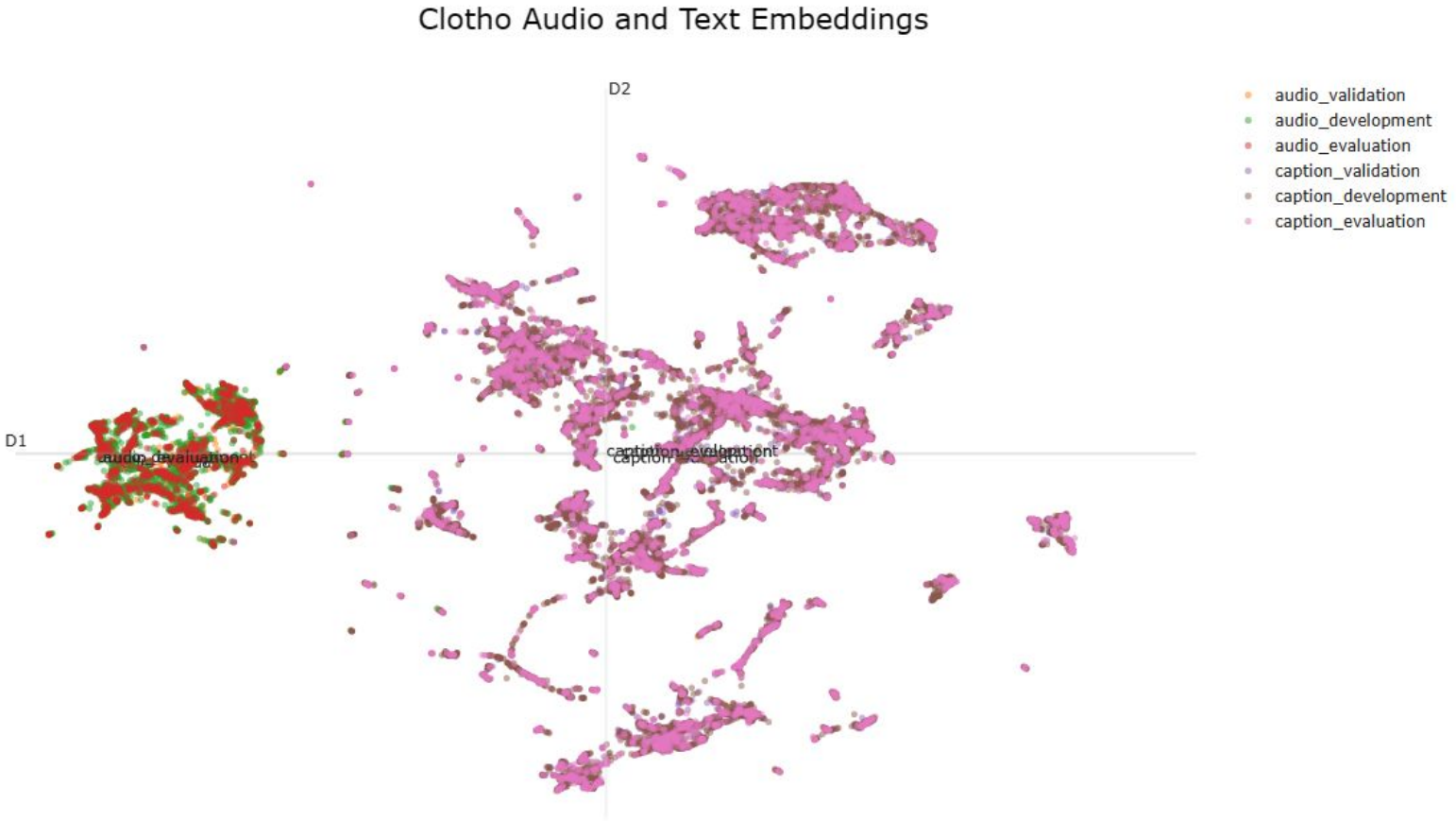
This audio sounds like

(Audio Prefix from Transformer Mapper)

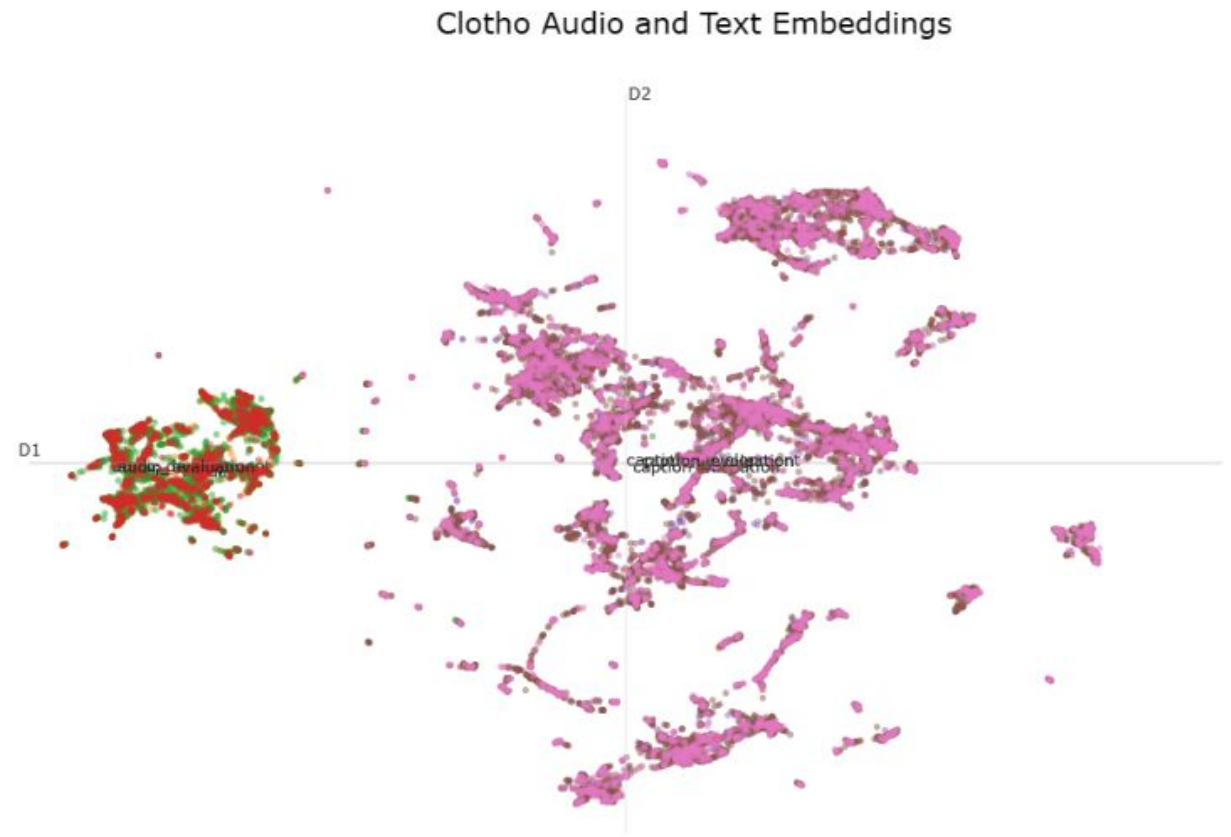
❄️ - Frozen

🔥 - Learnable

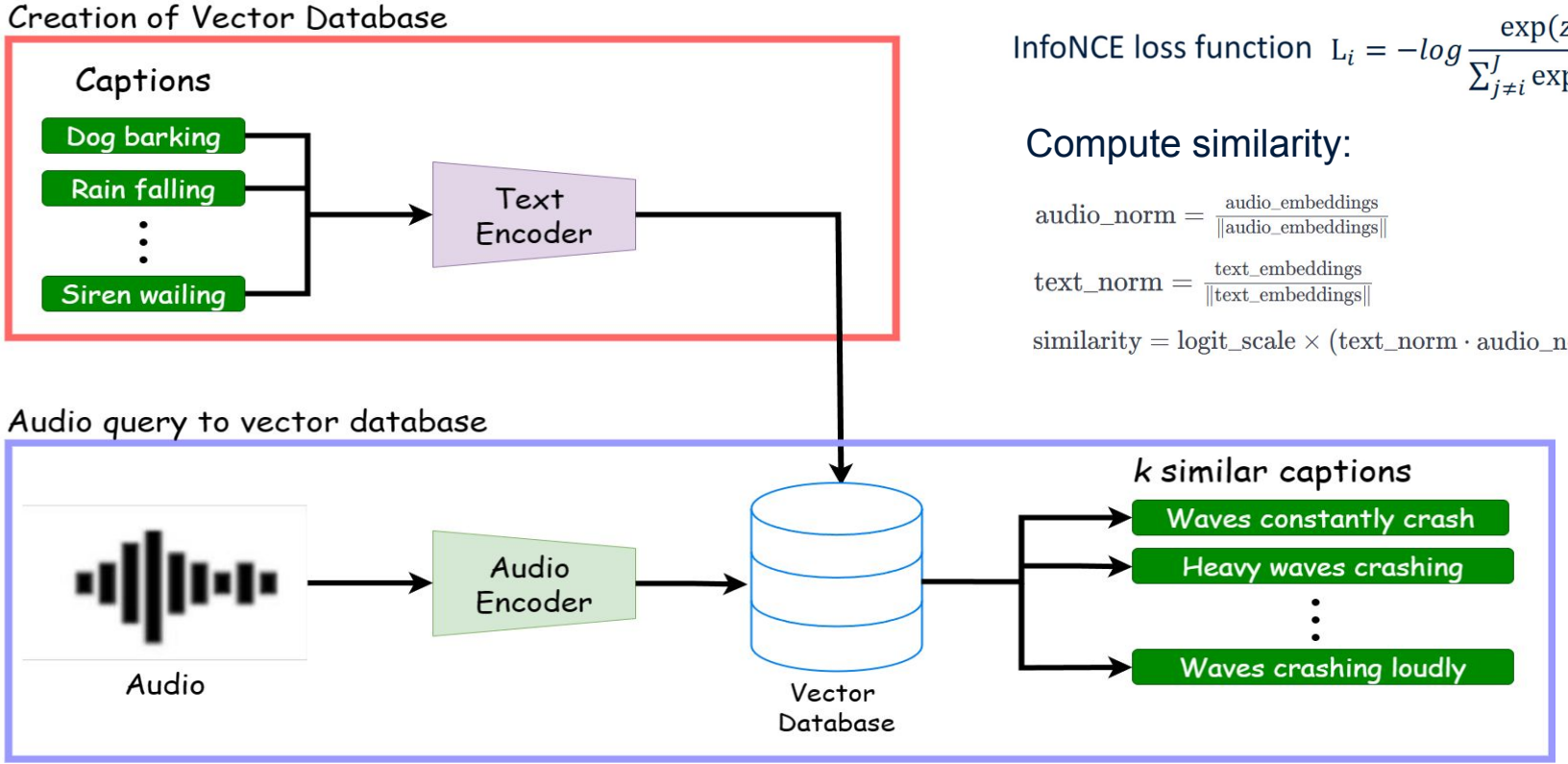
Audio and Text Embeddings MSCLAP



Audio and Text Embeddings MSCLAP



Vector DataBase for RAG



InfoNCE loss function $\mathrm{L}_i = -log \frac{\exp(z_i^T w_i/\tau)}{\sum_{j\neq i}^{J} \exp(z_i^T w_j/\tau)}$

Compute similarity:

$$\text{audio_norm} = \frac{\text{audio_embeddings}}{\|\text{audio_embeddings}\|}$$

$$\text{text_norm} = \frac{\text{text_embeddings}}{\|\text{text_embeddings}\|}$$

$$\text{similarity} = \text{logit_scale} \times (\text{text_norm} \cdot \text{audio_norm}^T)$$

Vector DataBase for RAG (MSCLAP)

Good: Heavy winds are blowing as waves from the ocean are crashing onto the shore. (original caption)



- Waves are crashing in along the coast as blustery winds cause interference.
- Waves crash loud and steady against the shore.
- Waves constantly crash loudly and with force against the shore.
- Waves constantly crash loudly and with force against shore.
- Waves crashing loudly into the shore multiple times.

Medium: A person repeatedly crushes a wet wrapper, making it smaller. (original caption)



- A bunch of paper cards filing together and hitting each other.
- A deck of cards is shuffled a few times.
- Distortion and shuffling as a person moves around and eventually dispenses tape from a roll.
- A sticky tape is being put together and ripped apart by someone over and over again.
- cards deck are being shuffled and pound on the table.

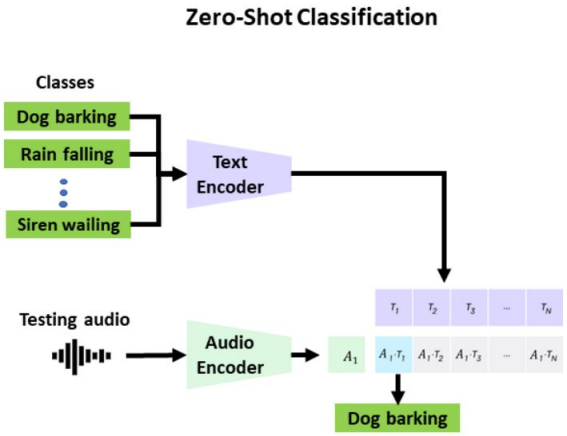
Vector DataBase for RAG (MSCLAP)

Good: Heavy winds are blowing as waves from the ocean are crashing onto the shore. (original caption)

- Waves are crashing in along the coast as blustery winds cause interference.
- Waves crash loud and steady against the shore.
- [...]
- [...]
- [...]

Medium: A person repeatedly crushes a wet wrapper, making it smaller. (original caption)

- A bunch of paper cards filing together and hitting each other.
- A deck of cards is shuffled a few times.
- [...]
- [...]
- [...]



Ontology

Semantic Role Labeling

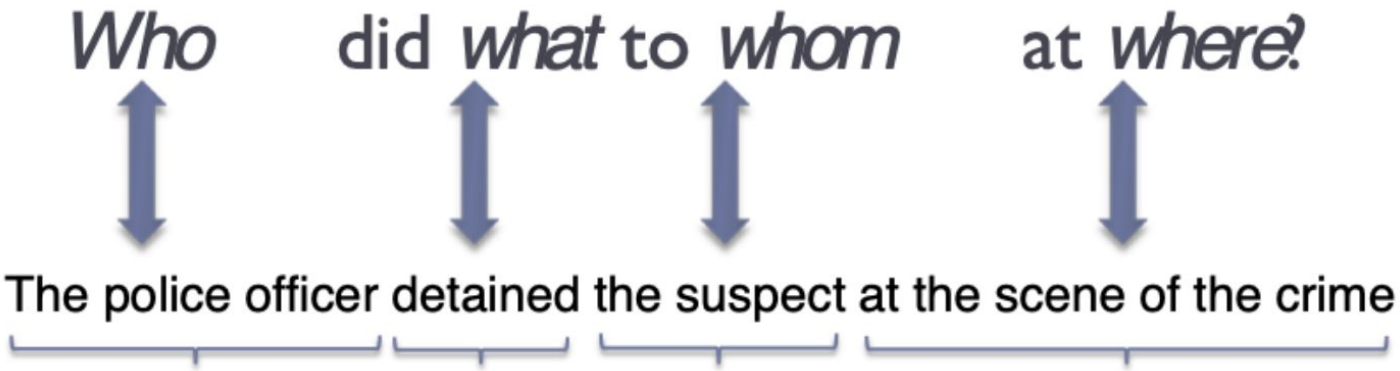


Image from: https://paperswithcode.com/task/semantic-role-labeling

Models for Semantic Role Labeling

- SRL mBERT base cased[1, 768]
- SRL XLMR large[1, 1024] / base[1, 768]

All models were fine-tuned on CoNLL-2012 data set

SRL Examples of captions

Caption: A plastic object is occasionally struck with a hard object while a piece of equipment is operated.

[A1: A plastic object -> AM-TMP: occasionally -> V: struck -> A2: with a hard object -> AM-TMP: while a piece of equipment is operated]

[A1: a piece of equipment -> V: operated]

SRL Examples of captions

Caption: A small baby making weird noises and the mother saying something.

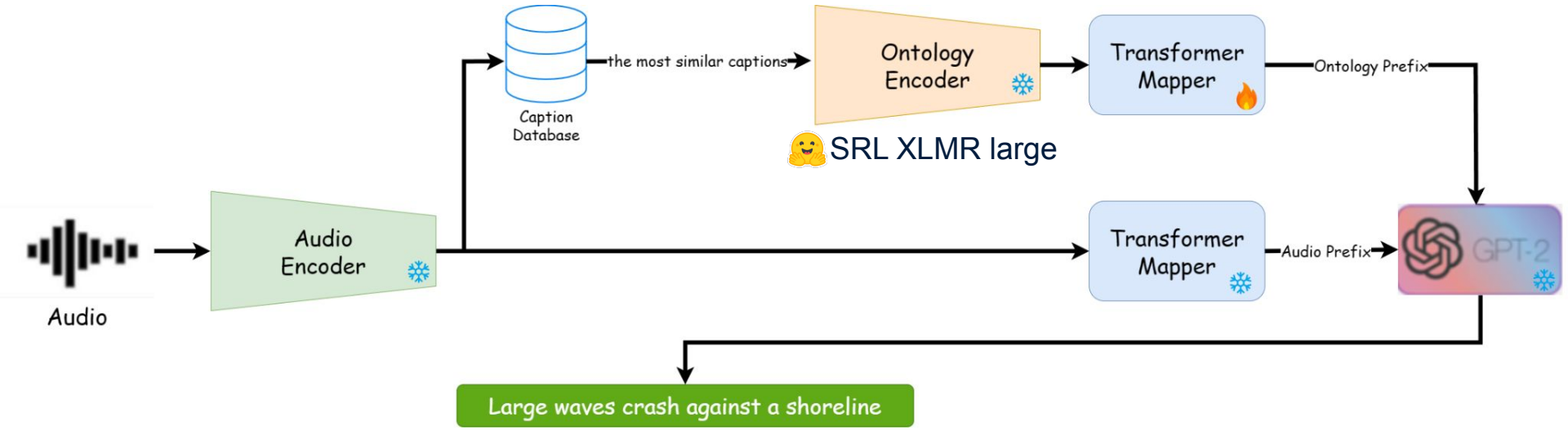
[A1: A small baby -> V: making -> A2: weird noises]

[A1: the mother -> V: saying -> A2: something]



Ontology based methods

Multimodality model



❄ - Frozen

🔥 - Learnable

Evaluation

Steps

1. **Audio Captioning Evaluation (AC Evaluation)**
   a. Find best Beam-Size
   b. Then find Temperature Value
   c. Now compare MSCLAP, MSCLAP RAG, and MSCLAP Multimodal

2. **Hallucination Detection**
   a. Find best metric and parameters
   b. Now compare MSCLAP, MSCLAP RAG, and MSCLAP Multimodal

AC Evaluation Metrics

- **Machine Translation metrics**: Sacre-BLEU, METEOR

  -> Linguistic Quality and Accuracy

- **Text Summarization**: ROUGE-L

  -> Conciseness and Relevance

- **Image captioning metrics**: SPICE, CIDER, SPIDEr

  -> Semantic Richness and Relevance

AC Evaluation Metrics

- Not AAC-specific metrics.
- But still the most commonly used ones in AAC related articles.

Best Beam-Size (MSCLAP)

		Sampling = False		
Metric		**Beam 1**	**Beam 5**	**Beam 10**
Sacrebleu		5.824	**8.992**	6.867
Meteor		0.428	0.408	**0.559**
Spice		0.108	0.087	**0.153**
Cider		0.475	0.317	**0.721**
Spider		0.292	0.202	**0.437**
Rouge-L	r	**0.030**	**0.030**	**0.030**
	p	**0.063**	0.056	0.056
	f	**0.041**	0.039	0.039

Best Temperature (MSCLAP)

		Beams 5		
Metric		**Temperature 0.5**	**Temperature 0.75**	**Temperature 1**
Sacrebleu		9.663	**9.896**	8.992
Meteor		0.486	**0.506**	0.352
Spice		**0.156**	0.154	0.087
Cider		0.545	**0.552**	0.317
Spider		0.350	**0.353**	0.202
Rouge-L	r	**0.030**	**0.030**	**0.030**
	p	**0.061**	0.059	0.056
	f	**0.040**	0.039	0.039

AC Evaluation (All Models)

Metric		MSCLAP	MSCLAP RAG	MSCLAP Multimodal
Sacrebleu		9.896	8.63	**10.959**
Meteor		**0.506**	0.321	0.359
Spice		**0.154**	0.085	0.086
Cider		**0.552**	0.256	0.263
Spider		**0.353**	0.171	0.174
Rouge-L	r	0.030	0.030	**0.031**
	p	**0.059**	0.054	0.052
	f	**0.039**	0.038	**0.039**

Fixed Params -> [beams:5, temp:0.75]

Hallucination detection

Models

- GPT-judge
- Truth-O-Meter
- SelfCheckGPT
- MENLI

Models

- GPT-judge
- Truth-O-Meter
- **SelfCheckGPT**
- **MENLI**

SelfCheckGPT

- 0 resource approach (no external database)
- Rank factuality of responses
- Blackbox
- Compares sampled answers

MENLI

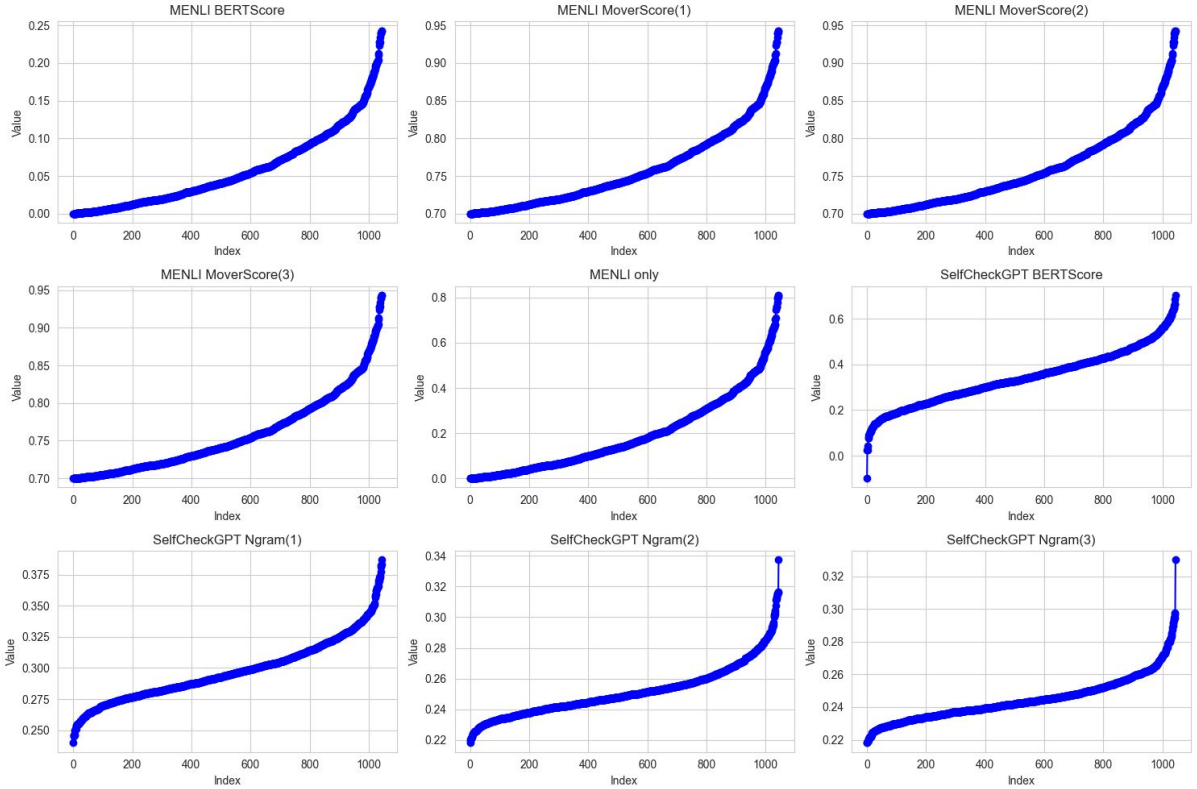
(Robust Evaluation Metrics from Natural Language Inference)

- BERT based, can combined with BERTScore or MoveScore
- Robust to adversarial attacks
- Penalize contradictions and checks factuality

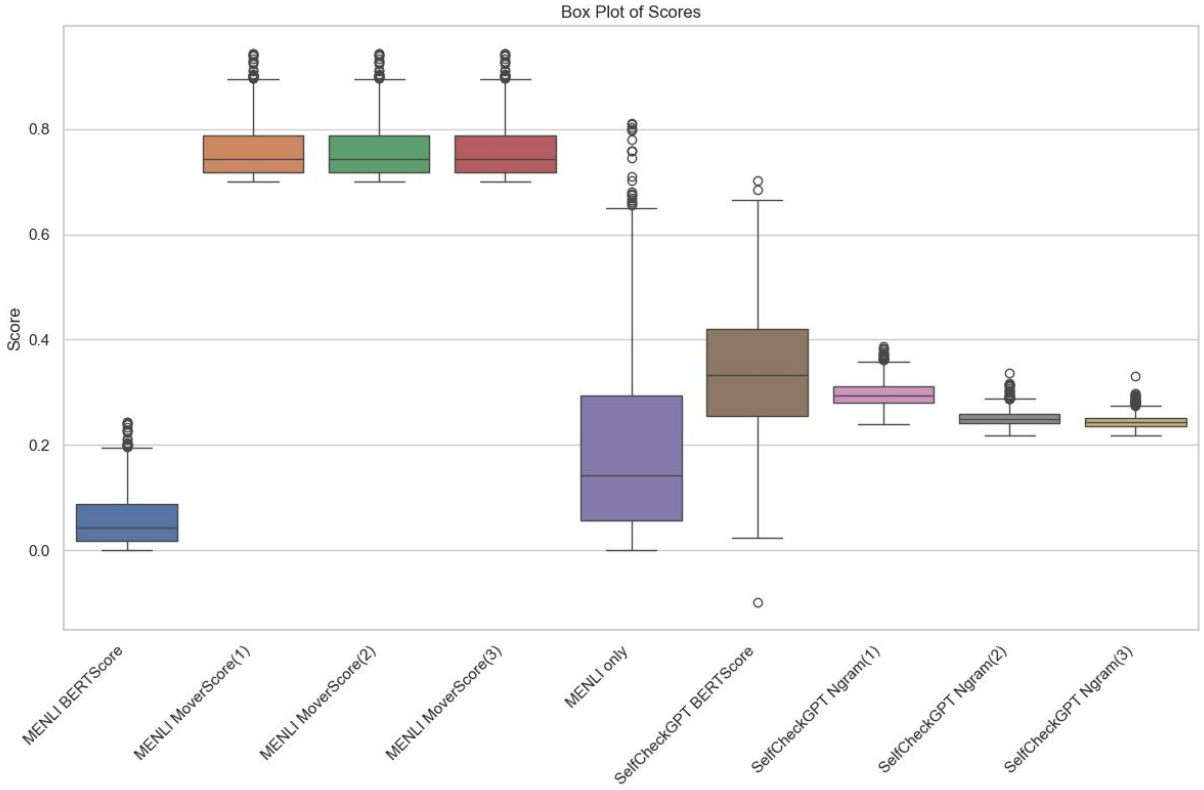
Hallucination detection

Metric Selection

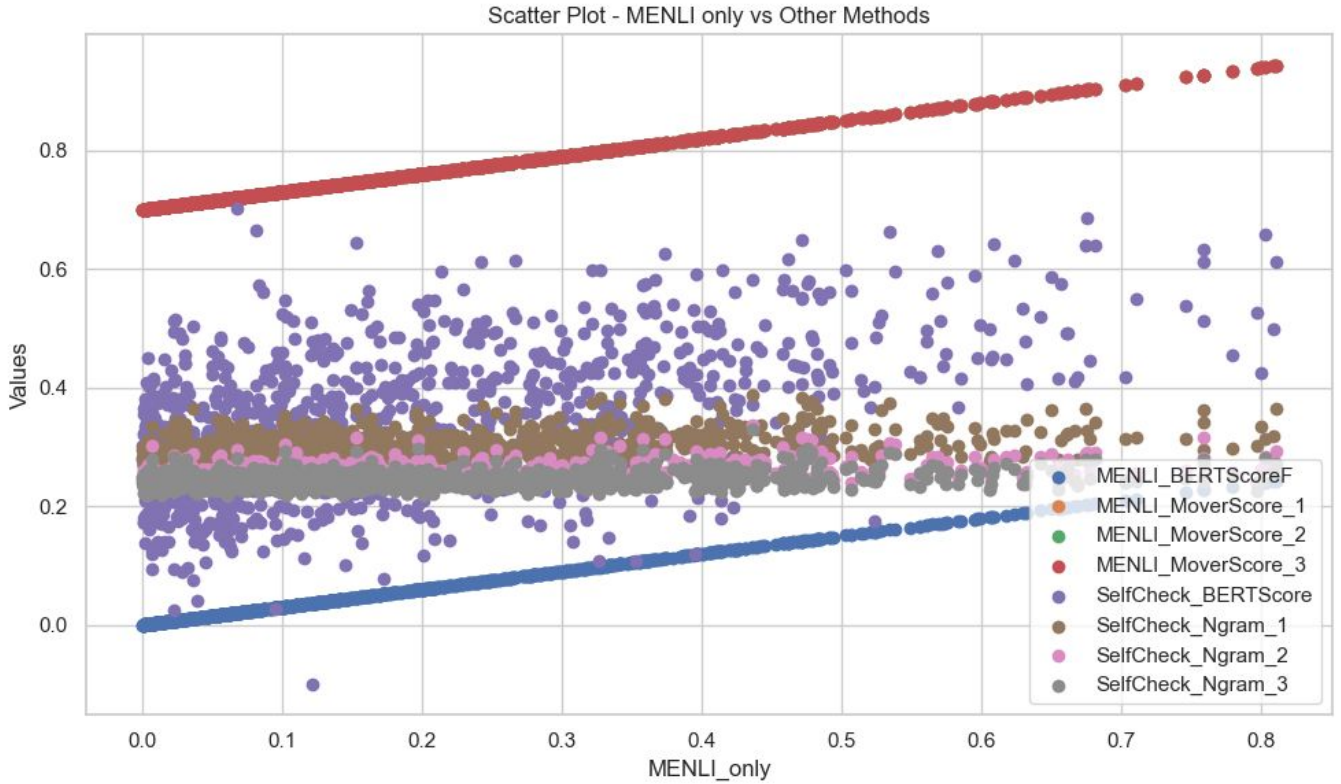
Metric selection



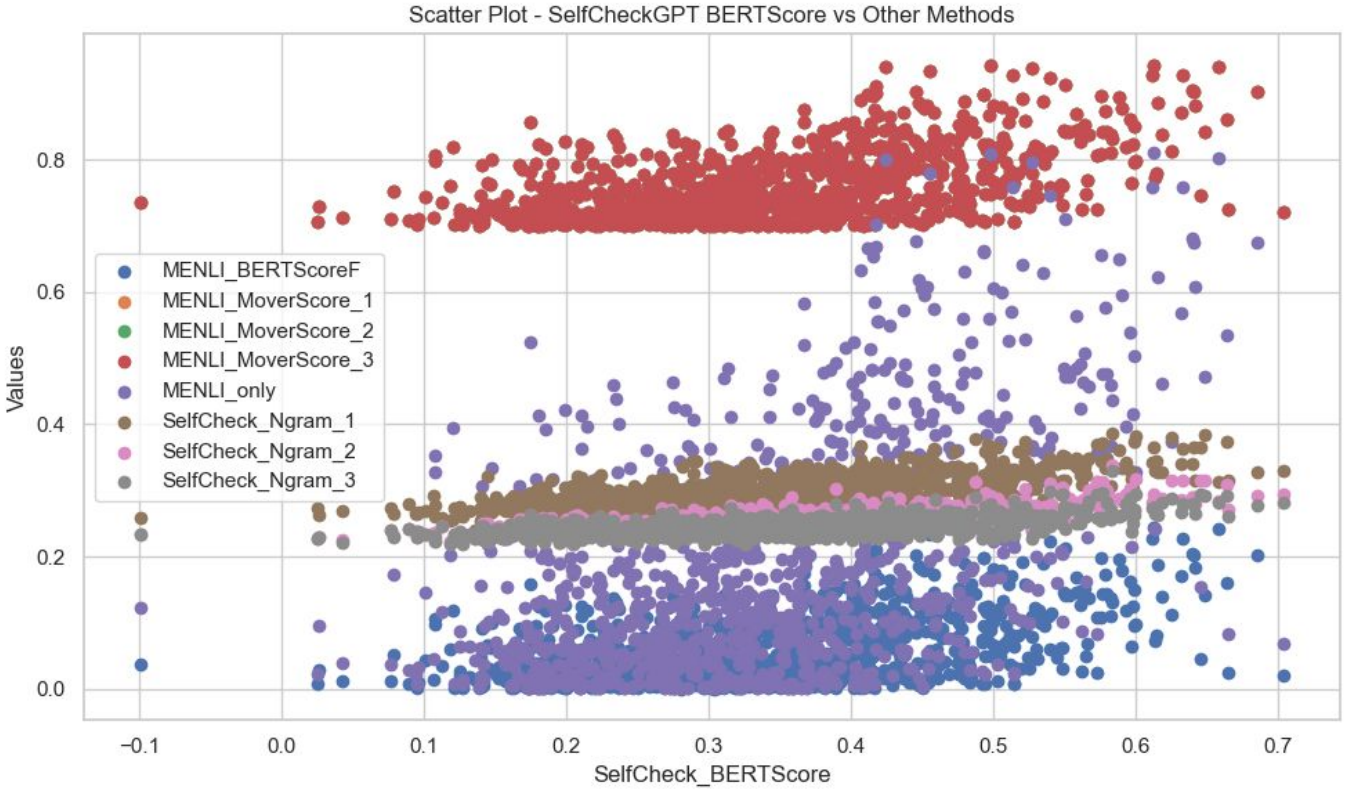
Metric selection



Metric selection



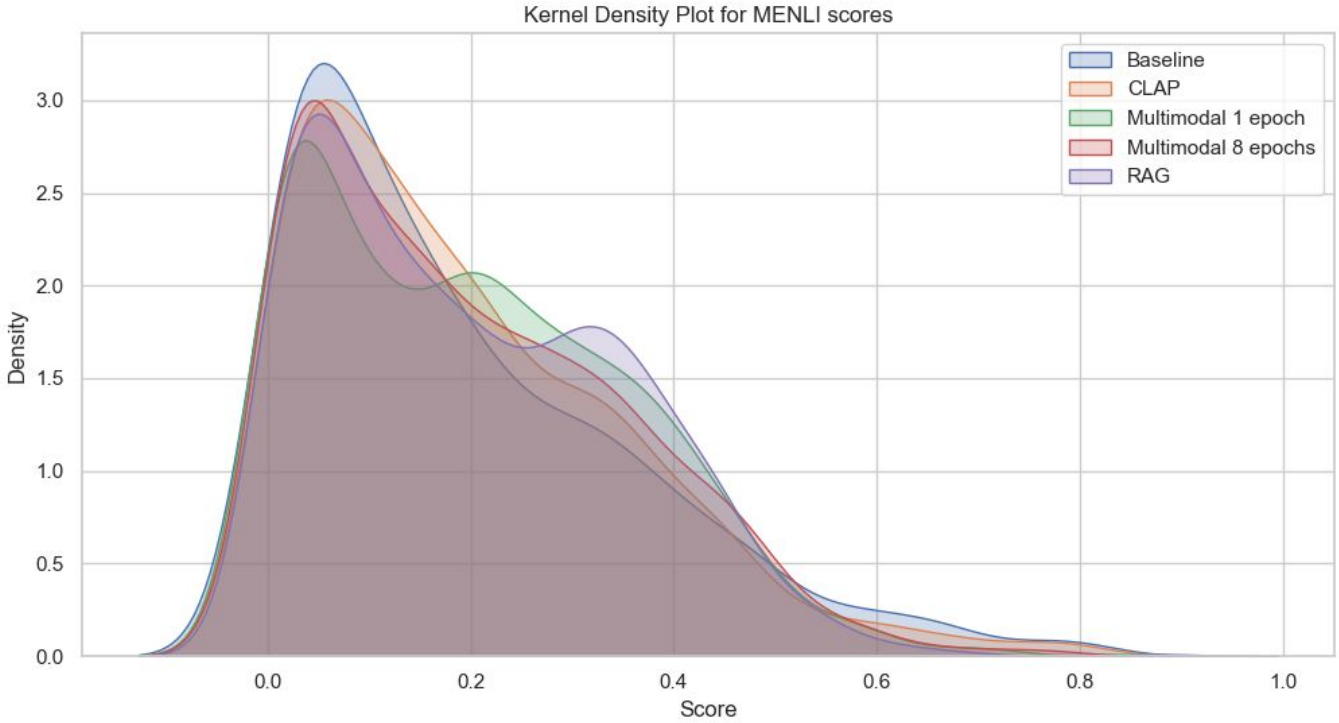
Metric selection



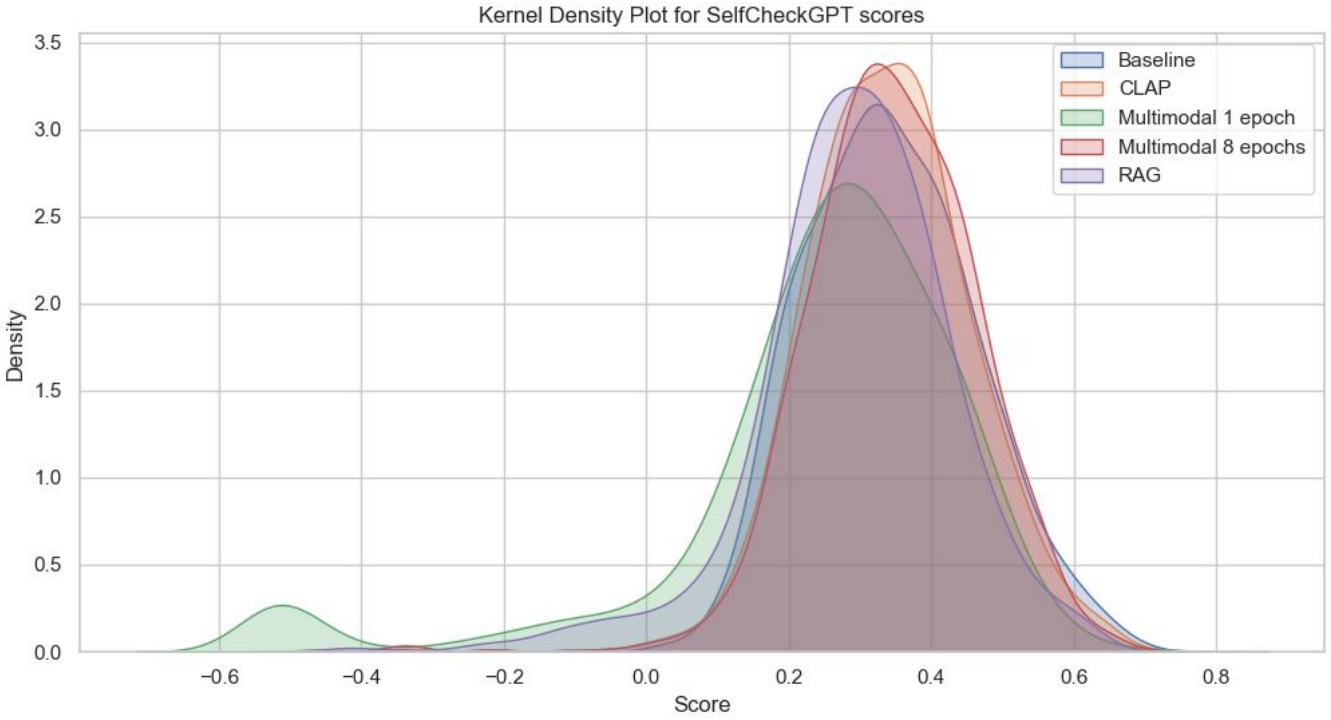
Hallucination detection

Model Evaluation

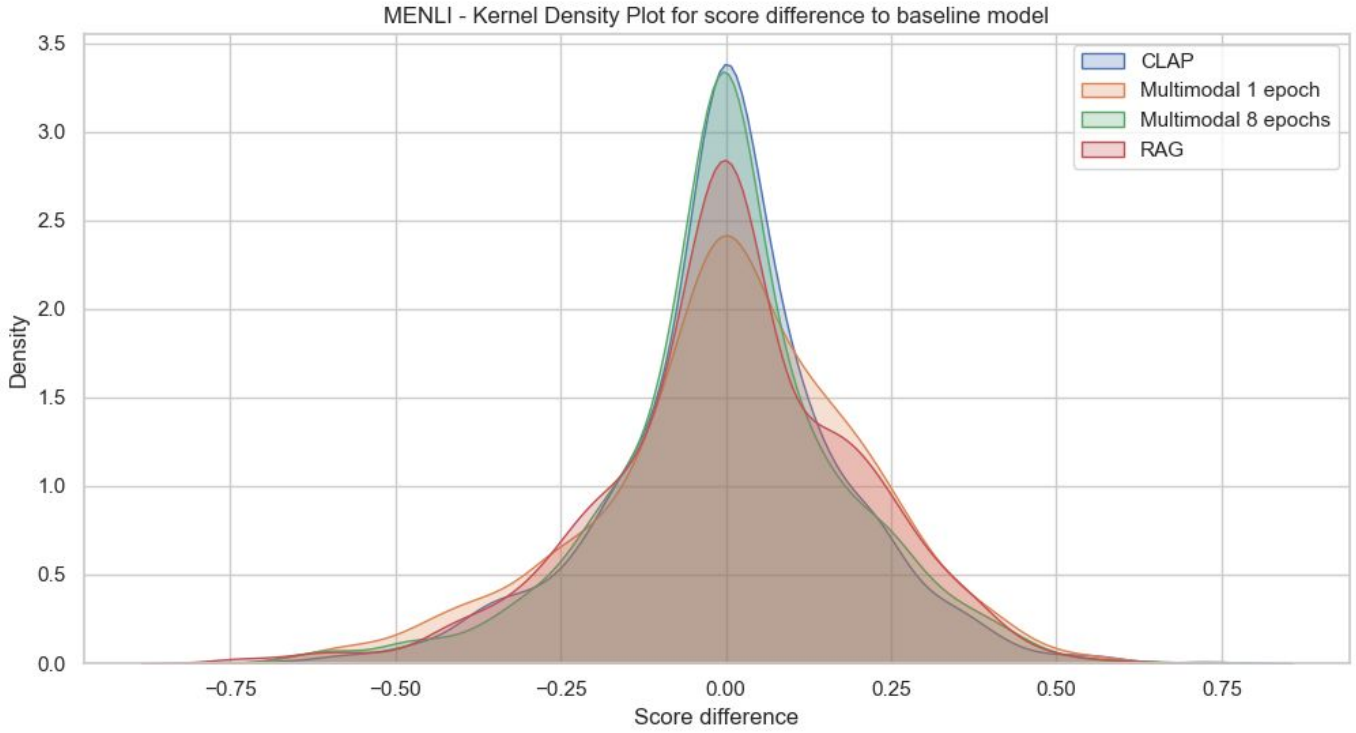
Model evaluation



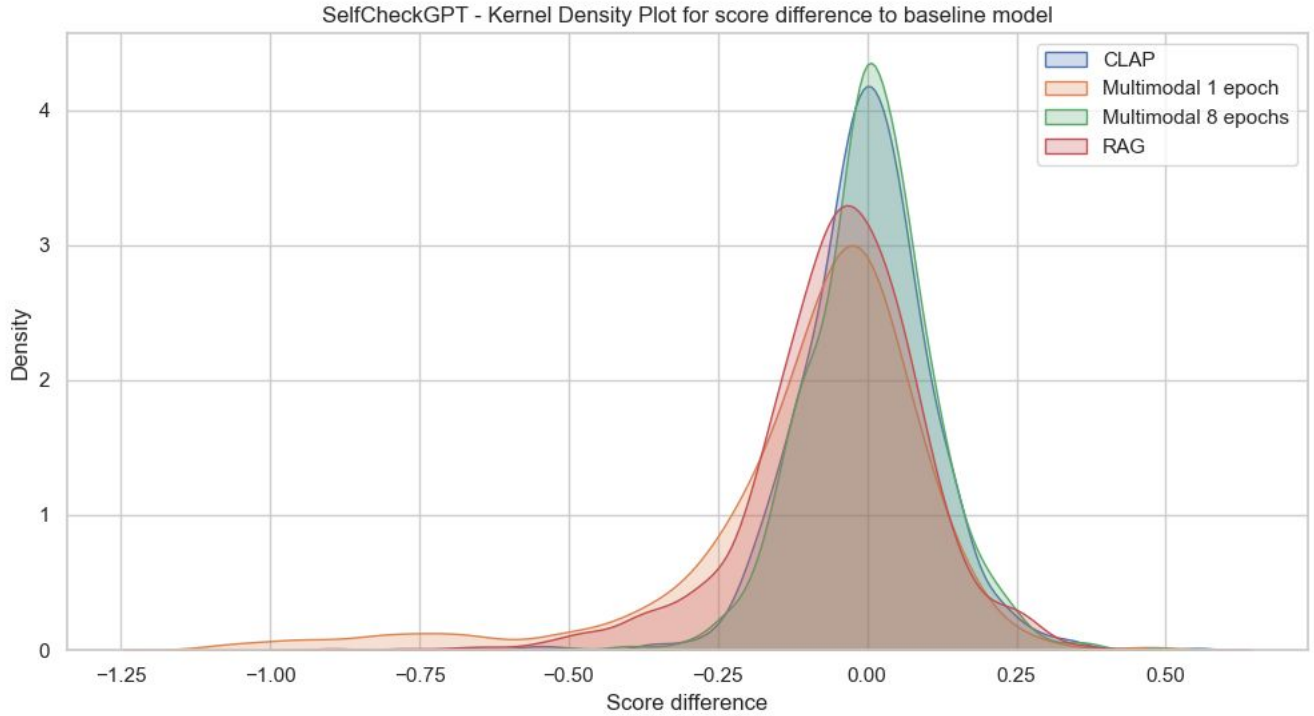
Model evaluation



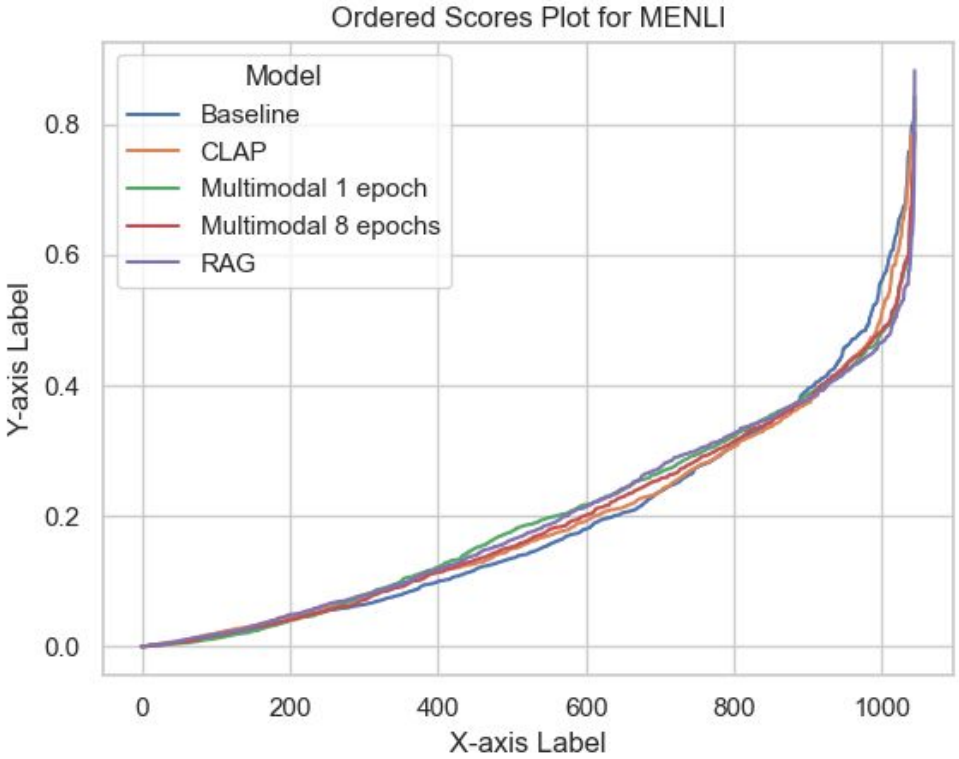
Model evaluation



Model evaluation



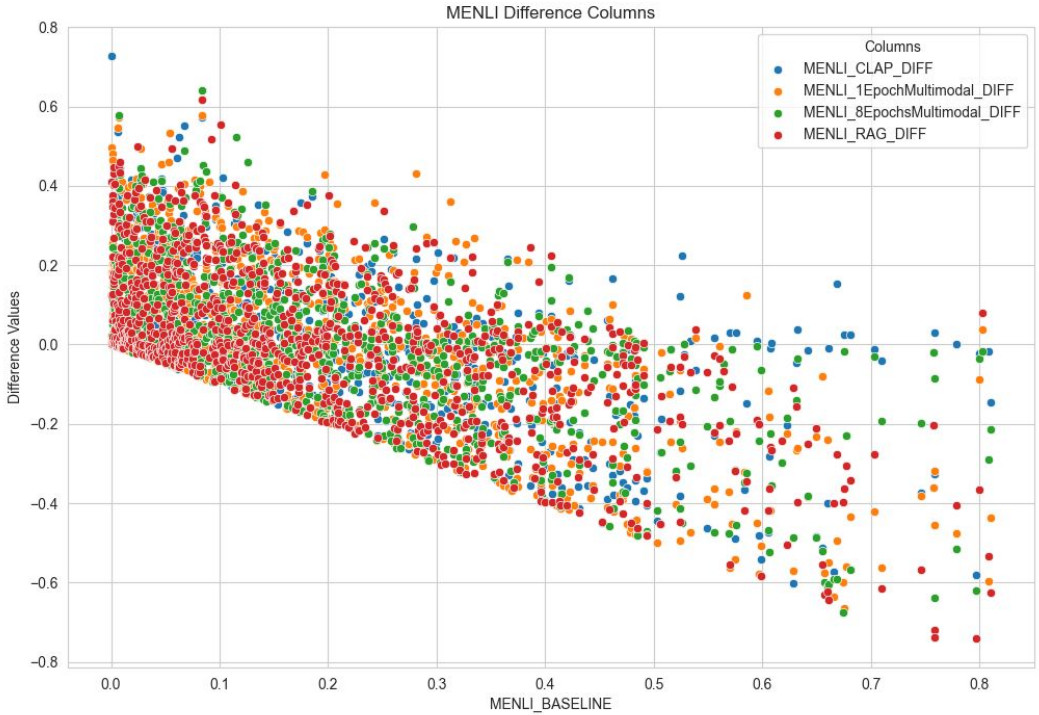
Model evaluation



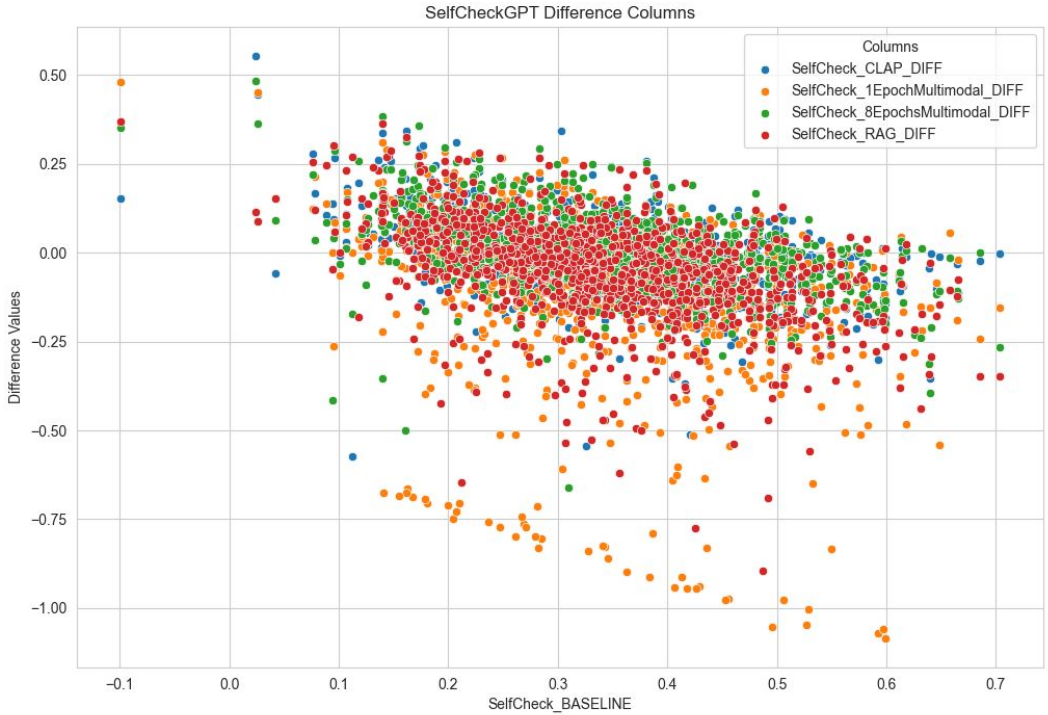
Model evaluation



Model evaluation



Model evaluation



Outputs observations

Outputs - Baseline vs CLAP compared with SelfCheckGPT

Baseline:
Coins are dropped onto the ground one at a time

CLAP:
A wind up wind up wind up wind up wind up wind up wind wind wind wind…

Score Difference: -0.57433

Outputs - Baseline vs CLAP compared with SelfCheckGPT

Baseline: An announcement is made over a public announcement system.



CLAP: A woman is talking over a loudspeaker while a crowd of people talk in the background.

Score Difference: +0.553404494451904

But this is not really hallucinated!

Outputs - Baseline vs CLAP compared with MENLI

Baseline: A motorcycle engine is started and then idles.



CLAP: A ball is bouncing in a steady rhythm.

Score Difference: -0.67281

CLAP is worse, but MENLI finds these hallucinations, but not Wind wind wind wind

Outputs - Baseline vs CLAP compared with MENLI

Baseline: Someone writing with a marker on a piece of paper.

CLAP: A metal object is being rubbed against a metal surface.

Score Difference: +0.5346036589364319

But here CLAP is better (compared to reference labels at least)

Outputs - Baseline vs 1 Epoch Multimodal

Some of the outputs are:
!!!

Outputs - Baseline vs 8 Epochs Multimodal compared with SelfCheckGPT

Bad score detected are of the type:

Water is bubbling and gurgling in a gurgling gurgling …

But also hallucinations! Such as:
A basketball is being bounced at a steady pace.



The model has clear weaknesses

Outputs - Baseline vs 8 Epochs Multimodal compared with MENLI



Baseline: Buzzing of an insect with birds chirping in the background.

New: A siren wails while birds chirp in the distance.

But MENLI locates some improvements

Outputs - Baseline vs RAG: quite mitigated results

Baseline: A dog is barking while a car drives by.



RAG: A motorcycle speeds by and then slows down as a car passes by.

Baseline: Buzzing of an insect with birds chirping in the background.



RAG: A siren going off with some birds chirping in the background.

Discussion

Discussion

- **MSCLAP RAG:**
  - Not good performance.
  - Pre-trained language model's poor ability to incorporate additional context.
  - This limitation is critical as retrieval captions which are used as context, plays a crucial role in describing audio content.
  - Needs larger and more diverse training dataset.

Discussion

- **MSCLAP Multimodal:**
  - Does not outperform the established baseline.
  - The retrieval-based captions, are incapable of transfering the full semantic meaning of the audio into the language model.
  - This means that we need a deeper analysis of the audio content at a semantic level.
  - Extraction of ontology embeddings directly from audio embeddings.

Discussion

- **Future Works?**
  - Other hallucination metrics.
  - Using a hallucination metric during fine-tuning.
  - Alternatives to CLAP.
  - Using more advanced knowledge graph.

Conclusions

Conclusions

Research Question 1:

What type of knowledge infusion contributes the most in reducing hallucinations and improving the accuracy and relevance of generated captions in our model?

- Multimodal shows the most promising results. Meaning, introducing additional context such as ontology as an embedding, leads to some improvement.

Conclusions

Research Question 2:

What multimodal method allows for an optimal fusion of our ontology model and audio encoder?

- Using an additional **Mapper** to pass embedding to Language Model allows us to introduce additional context and increase efficiency of the model.

Conclusions

Research Question 3:

What are the most robust and insightful evaluation metrics for assessing the quality, relevance, and factual correctness of the captions generated by the proposed model?

- For performance evaluation, Sacre-BLEU and METEOR.
- For hallucination detection, results are mitigated and would require additional investigation in alternative methods.

Thank you for Listening

