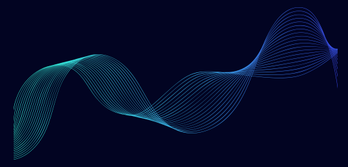
Mirror Minds: An Empirical Study on Detecting LLM-Generated Text via LLMs

Josh Baradia, Dept of CSE, PES University, Bengaluru, India
Shubham Gupta, and Suman Kundu, Dept of CSE, Indian Institute of Technology Jodhpur, India

COLING 25 Detecting AI Generated Content (GenAIDetect) - 19th January 2025



Presented by Josh Baradia
joshbaradia22@gmail.com

OVERVIEW

INTRODUCTION

- Large Language Models (LLMs) are increasingly integral in text generation, offering intelligent alternatives and gradually replacing traditional search engines.

INTRODUCTION

- Large Language Models (LLMs) are increasingly integral in text generation, offering intelligent alternatives and gradually replacing traditional search engines.

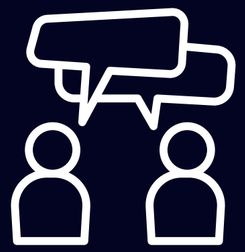
- LLMs are widely used for:



INTRODUCTION

- Large Language Models (LLMs) are increasingly integral in text generation, offering intelligent alternatives and gradually replacing traditional search engines.

- LLMs are widely used for:



- Their applications extend across various fields:




CHALLENGES

- Despite their remarkable capabilities, the misuse of LLMs is a growing concern:

CHALLENGES

- Despite their remarkable capabilities, the misuse of LLMs is a growing concern:

- Humans struggle to distinguish between AI-generated and human-written content, necessitating automated detection systems.

CHALLENGES

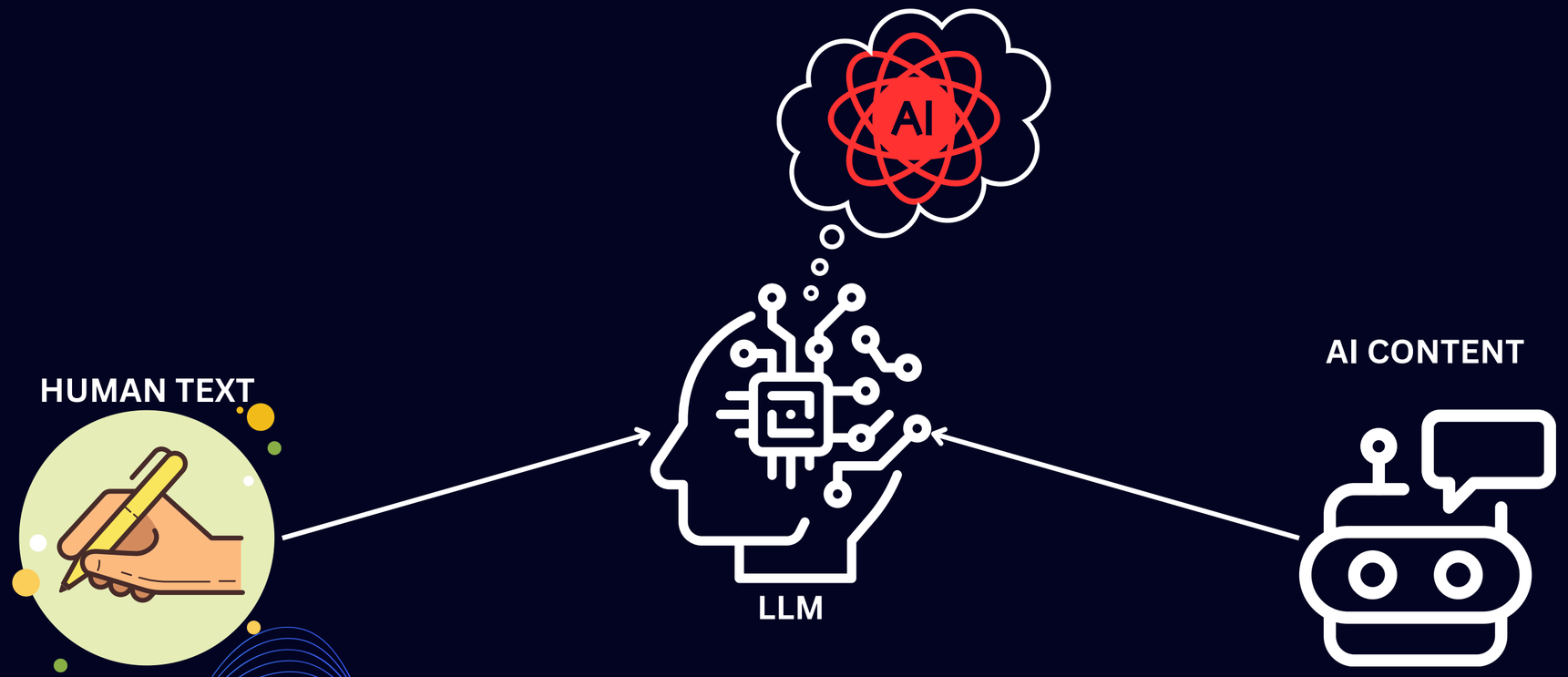
- Despite their remarkable capabilities, the misuse of LLMs is a growing concern:

- Humans struggle to distinguish between AI-generated and human-written content, necessitating automated detection systems.

- LLMs, with their pre-training and instruction tuning, bypass the need for task-specific learning, making their outputs more indistinguishable.

PROBLEM STATEMENT

"Can the generative capabilities of LLMs be utilized to detect AI-generated content?"



RELATED LITERATURE

Feature-Based Detection:

- *Rare bigram frequency*
- *N-gram frequencies*
- *Top-k word patterns*

Grechnikov et al., 2009; Badaskar et al., 2008; Gehrmann et al., 2019b
Guo et al. (2023); OpenAI, Jan 2023; Tian, 2023; Mitchell et al., 2023
Kirchenbauer et al. (2023); Yang et al., 2024

RELATED LITERATURE

Feature-Based Detection:

- *Rare bigram frequency*
- *N-gram frequencies*
- *Top-k word patterns*

Model-Based Detection:

- *Fine-tuned RoBERTa*
- *OpenAI Text Classifier And GPTZero*
- *DetectGPT*

Grechnikov et al., 2009; Badaskar et al., 2008; Gehrmann et al., 2019b
Guo et al. (2023); OpenAI, Jan 2023; Tian, 2023; Mitchell et al., 2023
Kirchenbauer et al. (2023); Yang et al., 2024

RELATED LITERATURE

Feature-Based Detection:

- *Rare bigram frequency*
- *N-gram frequencies*
- *Top-k word patterns*

Model-Based Detection:

- *Fine-tuned RoBERTa*
- *OpenAI Text Classifier and GPTZero*
- *DetectGPT*

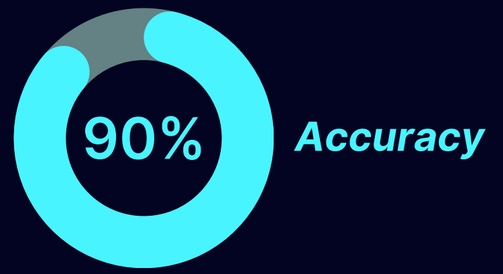
Alternative Approaches:

- *Watermarking*
- *DNA-GPT*

Grechnikov et al., 2009; Badaskar et al., 2008; Gehrmann et al., 2019b
Guo et al. (2023); OpenAI, Jan 2023; Tian, 2023; Mitchell et al., 2023
Kirchenbauer et al. (2023); Yang et al., 2024

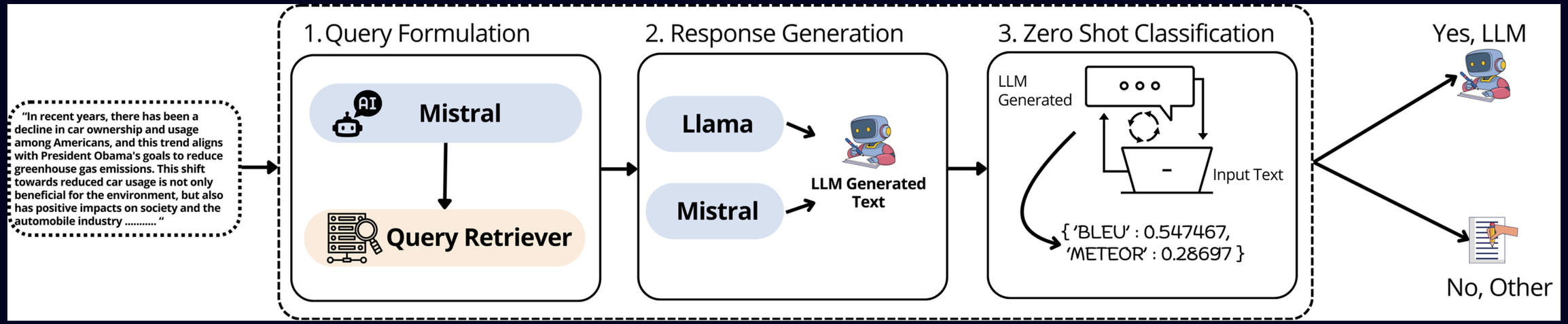
CONTRIBUTION

- Generate contextually relevant queries using LLMs in a zero-shot inference approach.
- Compare responses from multiple LLMs with original input text using BLEU and Meteor scoring.
- Comprehensive evaluation tested on publicly available four datasets.
- Prompt generation time ~2 seconds
- Achieved 90% accuracy in detecting AI-generated text.



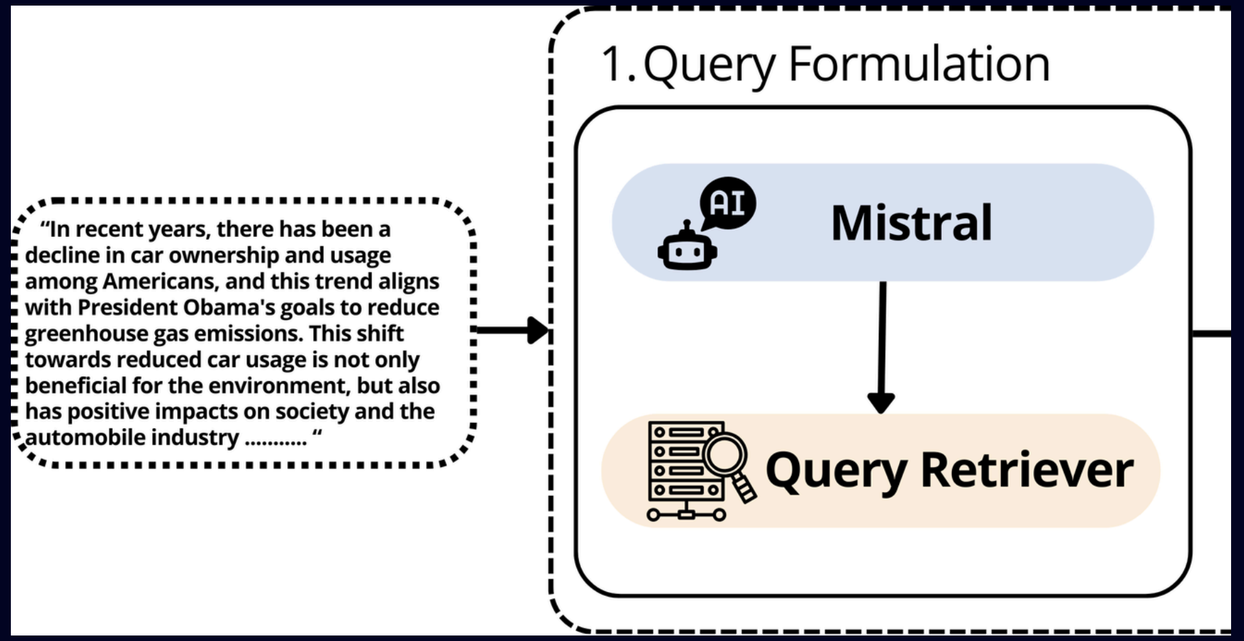
Kleczek, 2013(Revised Falcon-180B); Llama 70B; Muhammad, 2023; Shayan, 2023

OVERVIEW: ZERO-SHOT DETECTOR



STEP 1: QUERY FORMULATION

- **Key Role:**
  - *Crucial for zero-shot classification methodology*
  - *Interpret and extract query*
- **Importance**
- **Query, Q = Mistral(T),** T={w1, w2,⋯, wn}, consisting of a sequence of n tokens
- **Outcome**



LLM's Input Prompt:
Generate a query that encapsulates the main theme of the following text.{text}

STEP 1: EXAMPLE (DS1)

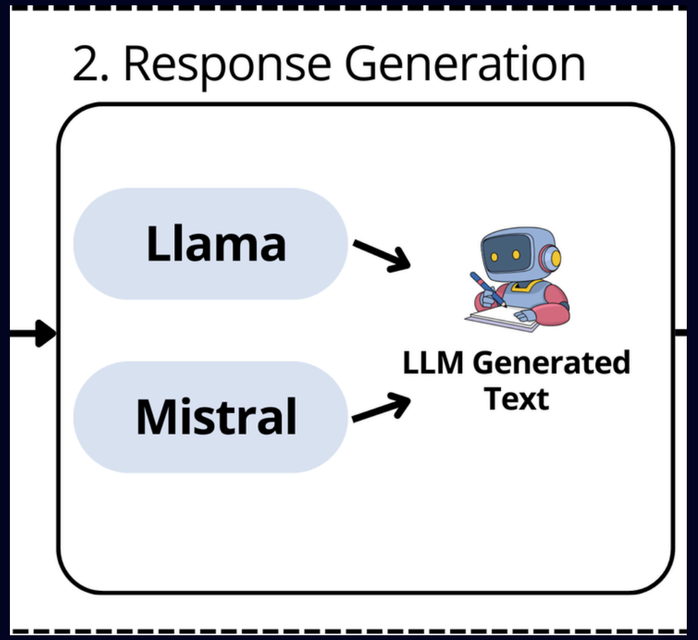
INPUT TEXT:(190 WORDS)

When people ask for advice, they sometimes talk to more than one person. This is because seaeking multiple opinions can help someone make a better choice. For example, if someone wants to buy a new phone.....always a good idea to seek multiple opinions before making a big decision.

QUERY:

What are the benefits of seeking multiple opinions before making a decision?

STEP 2: RESPONSE GENERATION



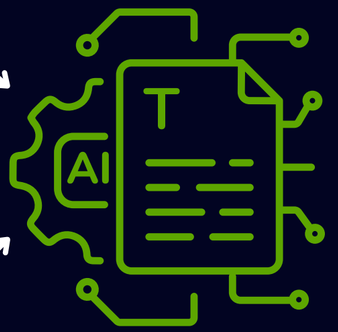
LLM's Input Prompt:
*Generate a response between {textWordCount} to {1.1*textWordCount} words for the following question.{Q}*

STEP 2: EXAMPLE (DS1)

Query:
What are the benefits of seeking multiple opinions before making a decision?

Llama Text:(197 words)
Seeking multiple opinions before making a decision can have numerous benefits. Firstly, it allows individuals to gain a more comprehensive understanding of the situation and the potential outcomes. By hearing from different perspectives, individuals can identify. . . . . .they value and respect those individuals. This can help build trust and strengthen relationships over time. Finally, seeking multiple opinions can help individuals develop a greater sense of confidence in their decision-making abilities. When individuals take the time to consider multiple perspectives.

Mistral Text:(183 words)
Seeking multiple opinions before making a decision is an essential practice that can bring numerous benefits. Here are some reasons why it's a good idea to gather various perspectives before finalizing a choice: 1. Broaden your understanding: Each person brings a unique perspective to be aware of any potential biases that may influence your thinking.....can help improve the quality of your decision-making. By considering various perspectives, you can weigh the pros and cons more thoroughly and make a more informed decision. 4. Enhance problem-solving skills: Seeking multiple opinions.

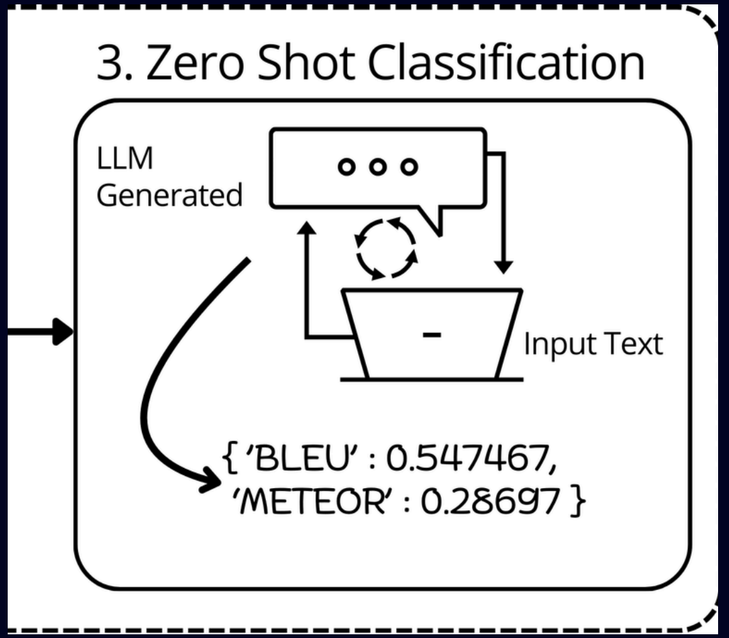


STEP 3: ZERO-SHOT CLASSIFICATION

Objective: To assess the similarity of generated texts T’ and T” to the reference input text T.

- Blue and Meteor quantify the closeness of the generated text to the original text in terms of syntax and semantics.
  - Bƒ= max(BLEU (T, T ′), BLEU (T, T ″))
  - Mƒ= max(Meteor(T, T ′), Meteor(T, T ″))

Originally Labelled as AI generated-BLEU=0.59, 0.53, METEOR=0.33, 0.28 for Llama3, Mistral respectively.



EXPERIMENTS

1.DATASET & EVALUATION METRICS

- **Threshold**
  - if BLEU score >= 0.52 then return 1
  - elif METEOR >= 0.27 then return 1
  - else return 0
  - Label:
    - 1 -> AI Generated text ;
    - 0 -> Human Text

Dataset	Description	Distribution
DS1	DAIGT dataset generated from Falcon-180B	1055 rows
DS2	DAIGT dataset generated from Llama-70B and Falcon-180B	7000 rows
DS3	Essay dataset generated by PaLM	1384 rows
DS4	Articles dataset consists of Human and AI generated text	20000 rows

Distribution of data across different datasets

1.DATASET & EVALUATION METRICS

- **Threshold**
  - if BLEU score >= 0.52 then return 1
  - elif METEOR >= 0.27 then return 1
  - else return 0
  - Label:
    - 1 -> AI Generated text ;
    - 0 -> Human Text
- **Thresholds determined using grid search:**
  - BLEU: Varying between 0.5 and 0.6.
  - Meteor: Varying between 0.2 and 0.3.

Dataset	Description	Distribution
DS1	DAIGT dataset generated from Falcon-180B	1055 rows
DS2	DAIGT dataset generated from Llama-70B and Falcon-180B	7000 rows
DS3	Essay dataset generated by PaLM	1384 rows
DS4	Articles dataset consists of Human and AI generated text	20000 rows

Distribution of data across different datasets

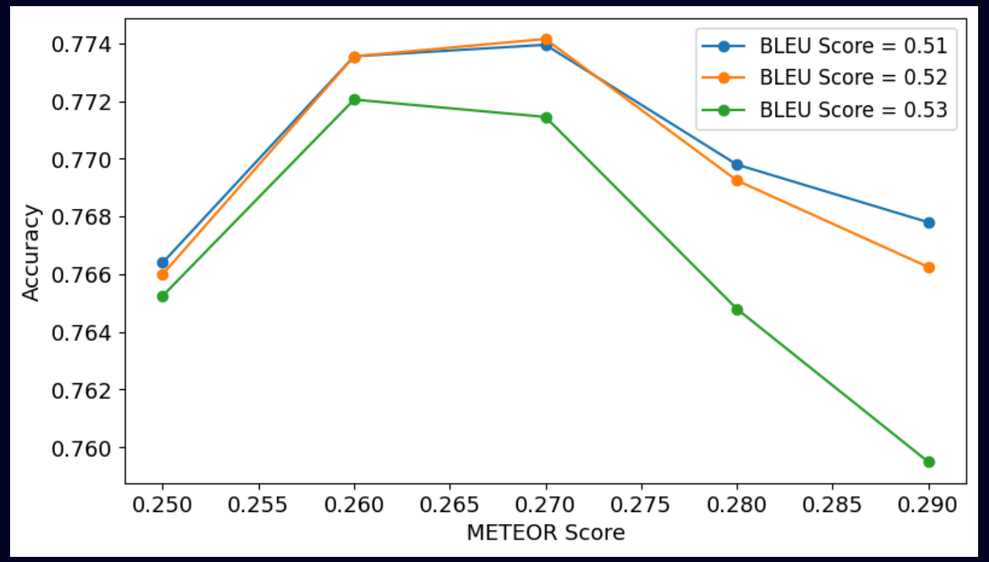
2. EVALUATION OF BLEU, METEOR & ACCURACY

BLEU Impact:

- BLEU 0.52 performs best in terms of accuracy.
- BLEU 0.53 leads to declining accuracy, demonstrating that excessively high BLEU scores may negatively impact performance.

Optimal METEOR Threshold:

- Accuracy peaks at METEOR = 0.27, across BLEU levels.



Comparison of accuracy among various BLEU and Meteor scores for DS4 dataset.

RESULTS

1. **Performance Across LLMs Individually**
2. **Improvement with Combined Models:**
   - Combining Llama3 and Mistral consistently yields better performance for BLEU, Meteor, and Accuracy across all datasets.
   - Indicates that integrating diverse LLMs leverages their complementary strengths.

Table 2: Performance comparison of different models across datasets

Dataset	Model	BLEU	METEOR	Accuracy
DS1	Ours-Llama	0.61	0.33	0.90
	Ours-Mistral	0.56	0.31	0.63
	Ours-(Llama+Mistral)	**0.62**	**0.34**	**0.92**
DS2	Ours-Llama	0.61	0.31	0.83
	Ours-Mistral	0.56	0.31	0.49
	Ours-(Llama+Mistral)	**0.62**	**0.32**	**0.86**
DS3	Ours-Llama	0.62	0.33	0.96
	Ours-Mistral	0.59	0.32	0.90
	Ours-(Llama+Mistral)	**0.63**	**0.34**	**0.98**
DS4	Ours-Llama	0.59	0.31	0.77
	Ours-Mistral	0.58	0.31	0.69
	Ours-(Llama+Mistral)	**0.60**	**0.32**	**0.78**

ABLATION STUDY

1. ROBUSTNESS ON DIFFERENT WORD RANGES

Samples were divided into the following word ranges:

- 0-100 words (detector struggles, insufficient contextual information)
- 100-200 words (richer contextual information, enhancing performance)

Dataset	Metric	0-100	100-200	200-300	300-400	400-500	500-600	>600
DS1	accuracy	0.50	0.82	0.90	0.96	0.89	-	-
	# rows	2	51	515	432	55	-	-
DS2	accuracy	0.0	0.41	0.70	0.83	0.91	0.90	0.89
	# rows	1	27	714	2161	2569	1208	320
DS3	accuracy	-	1.0	0.98	0.99	0.97	0.98	0.72
	# rows	-	1	180	658	474	60	11
DS4	accuracy	0.04	0.65	0.75	0.76	0.85	0.83	0.87
	# rows	84	1586	4790	6159	3790	1676	1915

Comparison of datasets across diffrent word ranges.

1. ROBUSTNESS ON DIFFERENT WORD RANGES

Samples were divided into the following word ranges:

- 0-100 words (detector struggles, insufficient contextual information)
- 100-200 words (richer contextual information, enhancing performance)
- 200-300 words (detection gets better as word count increase)
- 300-400 words
- 500-600 words
- Over 600 words

Dataset	Metric	0-100	100-200	200-300	300-400	400-500	500-600	>600
DS1	accuracy	0.50	0.82	0.90	0.96	0.89	-	-
	# rows	2	51	515	432	55	-	-
DS2	accuracy	0.0	0.41	0.70	0.83	0.91	0.90	0.89
	# rows	1	27	714	2161	2569	1208	320
DS3	accuracy	-	1.0	0.98	0.99	0.97	0.98	0.72
	# rows	-	1	180	658	474	60	11
DS4	accuracy	0.04	0.65	0.75	0.76	0.85	0.83	0.87
	# rows	84	1586	4790	6159	3790	1676	1915

Comparison of datasets across diffrent word ranges.

1. ROBUSTNESS ON DIFFERENT WORD RANGES

Samples were divided into the following word ranges:

- 0-100 words (detector struggles, insufficient contextual information)
- 100-200 words (richer contextual information, enhancing performance)
- 200-300 words (detection gets better as word count increase)
- 300-400 words
- 500-600 words
- Over 600 words

Dataset	Metric	0-100	100-200	200-300	300-400	400-500	500-600	>600
DS1	accuracy	0.50	0.82	0.90	0.96	0.89	-	-
	# rows	2	51	515	432	55	-	-
DS2	accuracy	0.0	0.41	0.70	0.83	0.91	0.90	0.89
	# rows	1	27	714	2161	2569	1208	320
DS3	accuracy	-	1.0	0.98	0.99	0.97	0.98	0.72
	# rows	-	1	180	658	474	60	11
DS4	accuracy	0.04	0.65	0.75	0.76	0.85	0.83	0.87
	# rows	84	1586	4790	6159	3790	1676	1915

Comparison of datasets across diffrent word ranges.

The detector's robustness improves with more contextual data, indicating its reliance on text length for effective zero-shot detection.

Case studies on short, long, and human-written texts provide deeper insights

EXAMPLE(DS4)

Sample(short length) of DS4 dataset-> Originally labeled as AI generated. Following are the BLEU & METEOR score of Llama3, Mistral.

BLEU score-> 0.37, 0.26 ; METEOR score-> 0.19, 0.16

Input Text:
Creating a new substance from chemicals is a complex process that requires extensive research. First, one must understand the chemical composition of different substances and learn the specific names of chemical compounds. Once knowledge is attained, experiments can be conducted in a lab to test different chemical reactions. The benefits of trial and error far outweigh the risks, and the potential for creating something incredibly valuable is infinite.

Query:
What are the steps involved in creating a new substance from chemicals and why is it a valuable process?

Llama Text:
The process of creating a new substance from chemicals involves several steps. First, chemists identify the desired properties of the new substance, such as its melting point, boiling point, and solubility. Next, they select the appropriate chemicals and react them.

Mistral Text:
Creating a new substance from chemicals, also known as organic synthesis or chemical synthesis, is a valuable process due to its potential to yield novel compounds with unique properties. The steps involved include: 1. Identifying the target.

3. INFERENCE TIME FOR DETECTION (2 CORE PHASES)

Dataset	QET	RGT	
		Llama3	Mistral
DS1	1.95	13.38	10.03
DS2	1.15	19.83	15.63
DS3	1.18	17.64	13.48
DS4	1.81	17.11	12.72

Analysis of time (in sec) taken across different phases 1) Query Formulation Time (QFT) and 2) Re-sponse generation Time (RGT)in zero shot detector.

CONCLUSION

- **Key Findings:**
  - Introduced a simple yet effective zero-shot detection method for identifying AI-generated text.
- **Methodology:**
  - Leverages the knowledge from LLMs' pre-training to detect their own generated outputs.
  - Classification via zero-shot inference without additional training.
  - Extracts input context and compares with LLM-generated responses to identify AI-generated text.
- **Evaluation:**
  - Tested on four publicly available datasets, including both in-domain and out-of-domain (OOD) data. Achieved effective detection of LLM-produced texts.
- **Key Contributions:**
  - Instruction tuning enhances alignment with user-expected responses in text detection tasks.

FUTURE SCOPE

- Broadening model scope
- Improved evaluation metrics
- Model optimization and Generalization
- Improved detection of short length text
- Practical application on smart devices like mobile phones, laptops etc.

Thank You!

Any Questions

Presented by Josh Baradia
joshbaradia22@gmail.com