Deciphering Textual Authenticity: A Generalized Strategy through the Lens of Large Language Semantics for Detecting Human vs. Machine-Generated Text

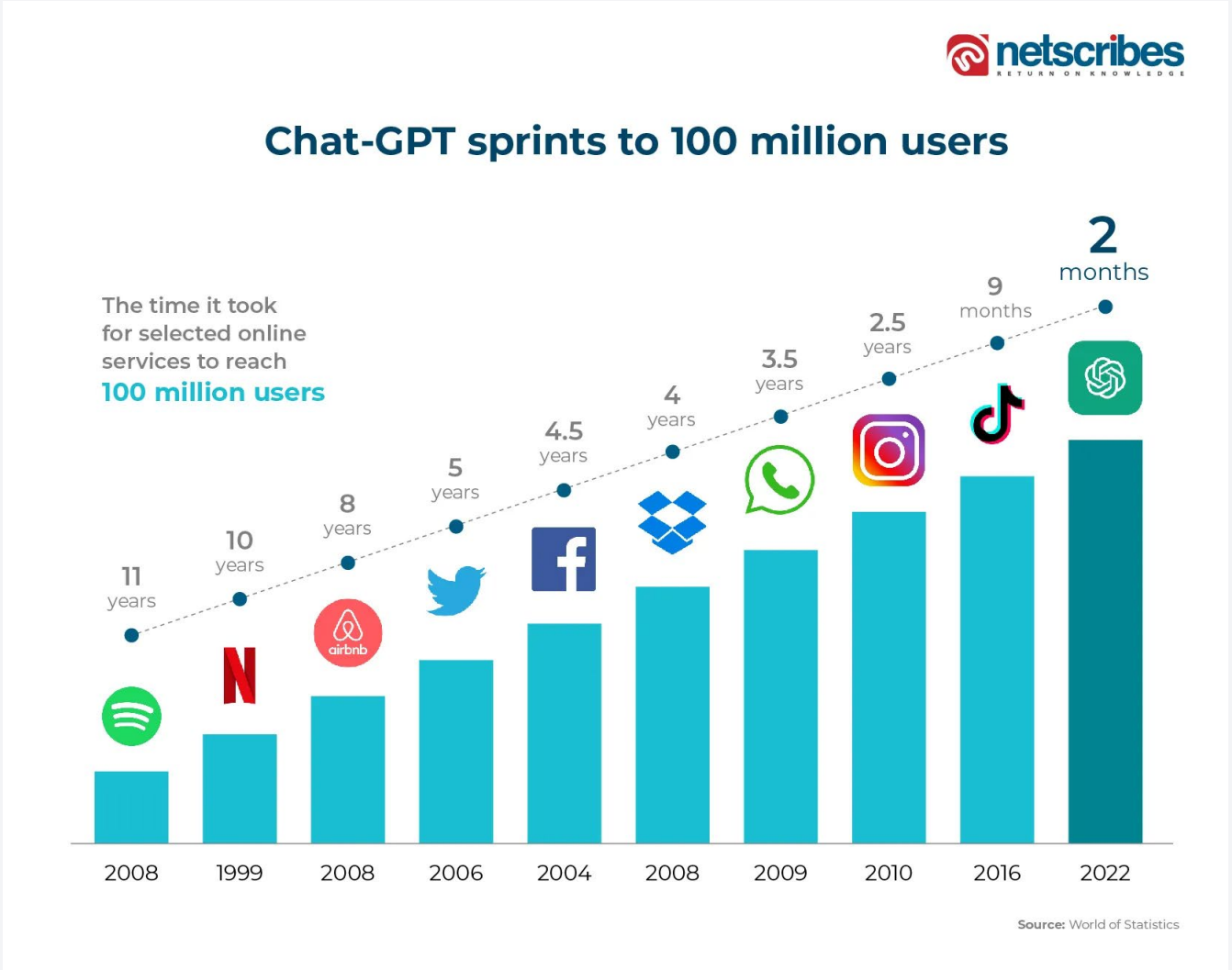
Authors: Mazal Bethany[1,2], Brandon Wherry[1,2,3], Emet Bethany[1,2,3],
Nishant Vishwamitra[2], Anthony Rios[2], and Paul Rad[1,2]

[1]Secure AI and Autonomy Lab, [2]University of Texas at San Antonio
[3]Peraton Labs

USENIX Security 2024

LLM Adoption

- LLMs are the fastest adopted tools in human history



Source: https://ai.plainenglish.io/chat-gpt-achieving-100-million-users-in-just-2-month-a-deep-analysis-a453e6f85acf

Machine Generated Text

- LLMs are revolutionizing text generation
- Growing concern:
  - Prevention of Misinformation
  - Intellectual Property Rights
  - Ethical Considerations
  - Detection of Malicious Activity (LLM phishing)
  - Incident Forensics
  - …

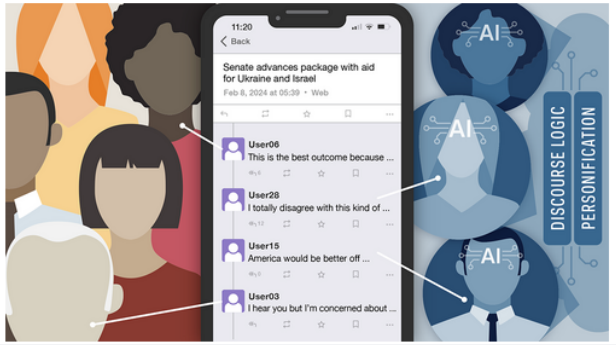
AI among us: Social media users struggle to identify AI bots during political discourse

February 27, 2024
Brandi Wampler

Share:

Artificial intelligence bots have already permeated social media. But can users tell who is human and who is not?

Researchers at the University of Notre Dame conducted a study using AI bots based on large language models — a type of AI developed for language understanding and text generation — and asked human and AI bot participants to engage in political discourse on a customized and self-hosted instance of Mastodon, a social networking platform.



Source: https://news.nd.edu/news/ai-among-us-social-media-users-struggle-to-identify-ai-bots-during-political-discourse/

Many LLMs and Text Domains

INDUSTRY

7 LLM use cases and applications in 2024

Learn about the top LLM use cases and applications in 2024 for streamlining business operations, automating mundane tasks, and tackling challenges.

With 700,000 Large Language Models (LLMs) On Hugging Face Already, Where Is The Future of Artificial Intelligence AI Headed?

By **Tanya Malhotra** - June 15, 2024

Large language models (LLMs) demonstrated higher error rates compared to humans in a clinical oncology question bank

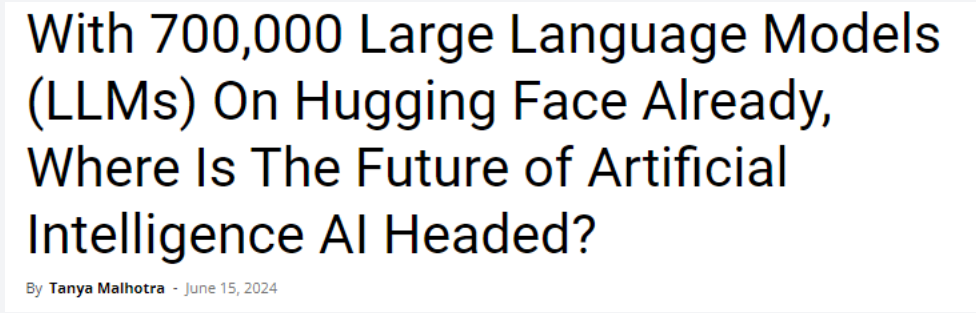
Claude

BY ANTHROP\C

cohere

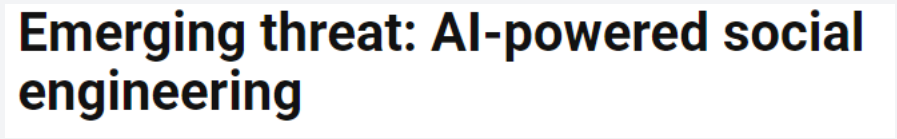
Emerging threat: AI-powered social engineering

Many LLMs and Text Domains



Urgent need:

Robust tools to detect machine-generated text across different LLMs and domains

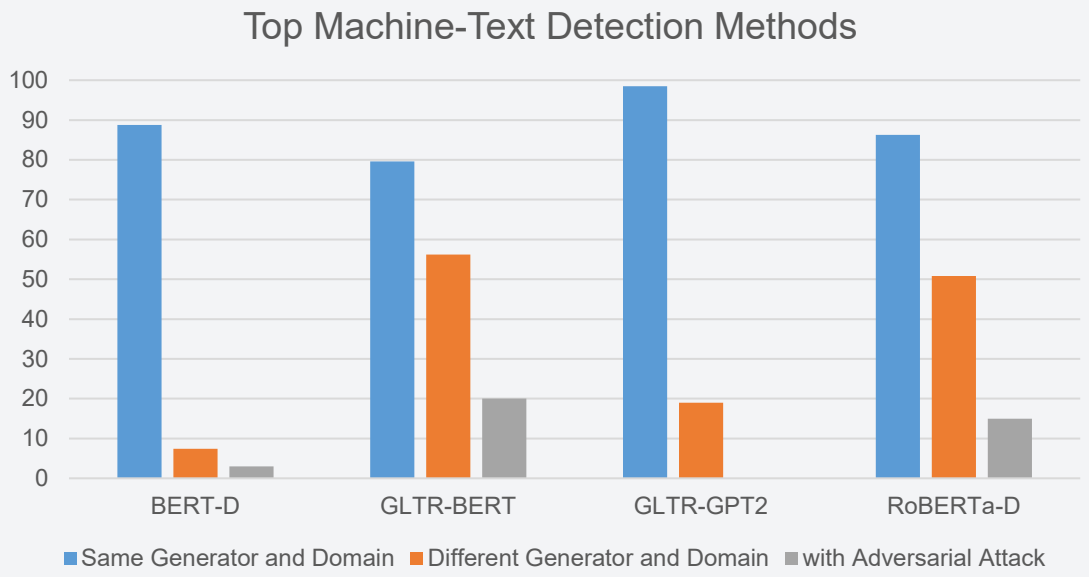


Dataset

- M4 dataset: 126,950 text samples
- Human sources: Wikipedia, Reddit, WikiHow, PeerRead, Arxiv
- Machine generators: ChatGPT, davinci-003, Cohere, Dolly, BLOOMz
- Additional in-the-wild datasets for real-world testing

Motivation Experiment: Studying Generalizability

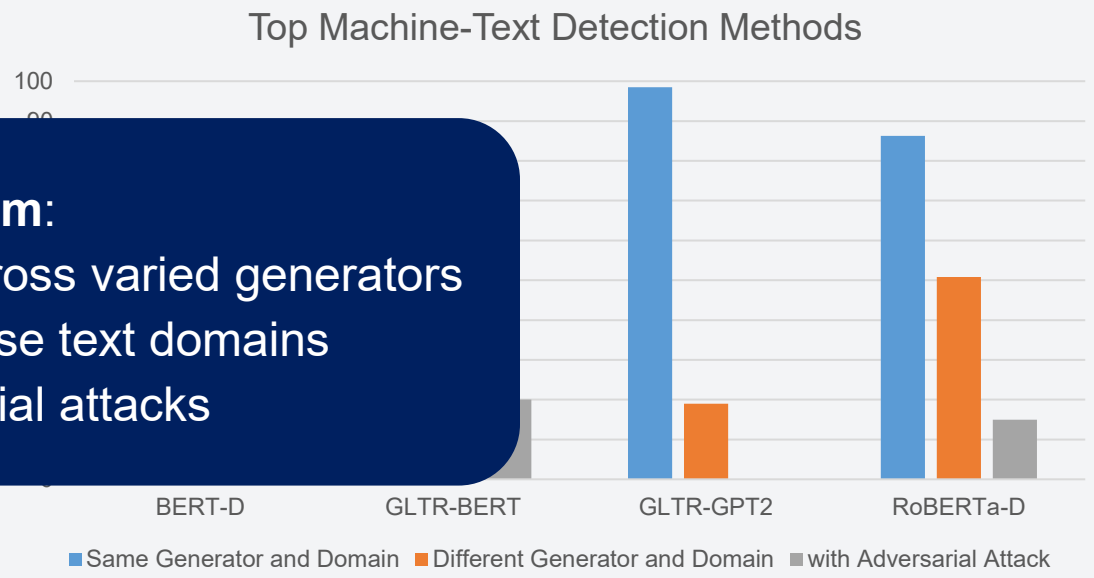
- Top detectors perform well when testing on the same text-generators and domain that they were trained on
- This is not a realistic real-world scenario
  - Too many LLMs and text domains!



Motivation Experiment: Studying Generalizability

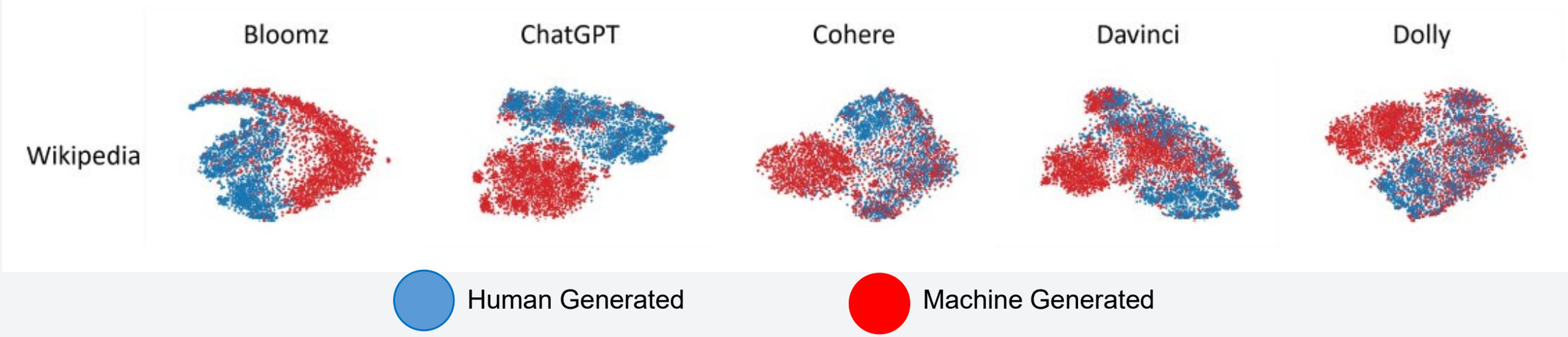
- Top detectors perform well when testing on the same text-generators an were trained c
- This is not a r scenario
  - Too many LLMs and text domains!

Problem:
- Limited performance across varied generators
- Poor adaptation to diverse text domains
- Vulnerability to adversarial attacks



Motivation Experiment: LLM Embeddings to Discern Text Origin

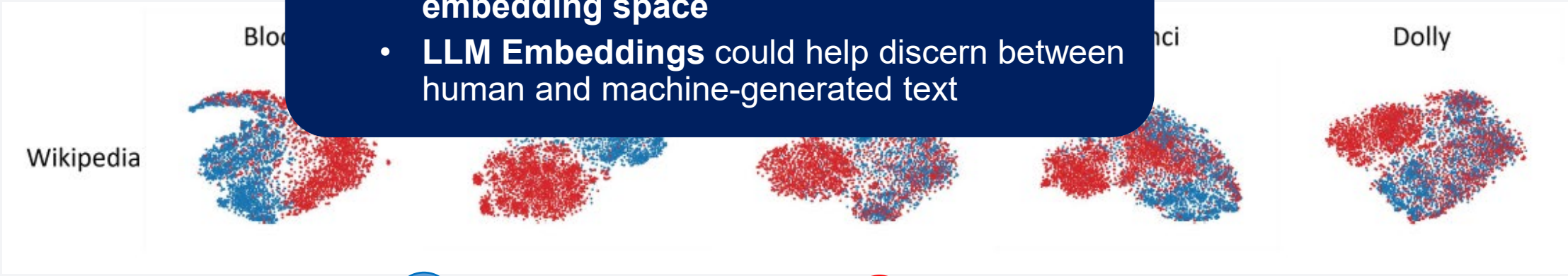
- LLMs have a great understanding of nuance in text
- Separation between human and machine texts on LLM embeddings (unsupervised)



Motivation Experiment: LLM Embeddings to Discern Text Origin

- LLMs have a great understanding of nuance in text
- Separation between human and machine texts on LLM embeddings (u

- **There is some separation between human and machine-generated text in the embedding space**
- **LLM Embeddings** could help discern between human and machine-generated text



Study Design

RQ 1: Can the encoders from LLMs be used to underpin an approach for generalized machine-generated text detection?

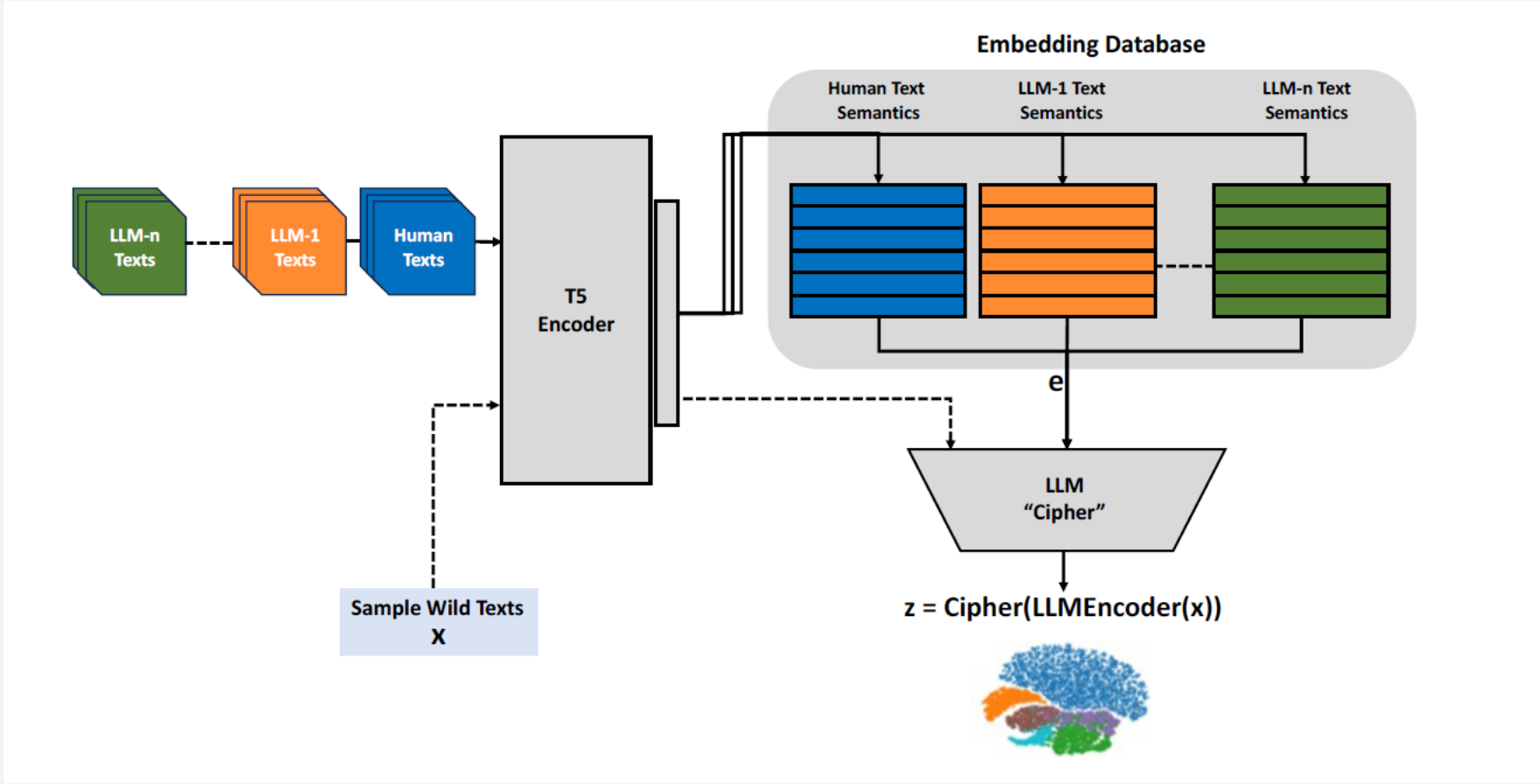
RQ 2: Are there fingerprints that can be detected to distinguish between multiple text generators?

Threat Model

- Two key parties
  - Adversaries producing machine-generated text
  - Stakeholders detecting machine-generated text
- Adversary capabilities
  - Can use public models, fine-tune existing models, or train new LLMs
  - Employ state-of-the-art LLMs to mimic human text
  - Can use adversarial attacks to evade detection
- Detector constraints: No prior knowledge of specific LLM or domain used

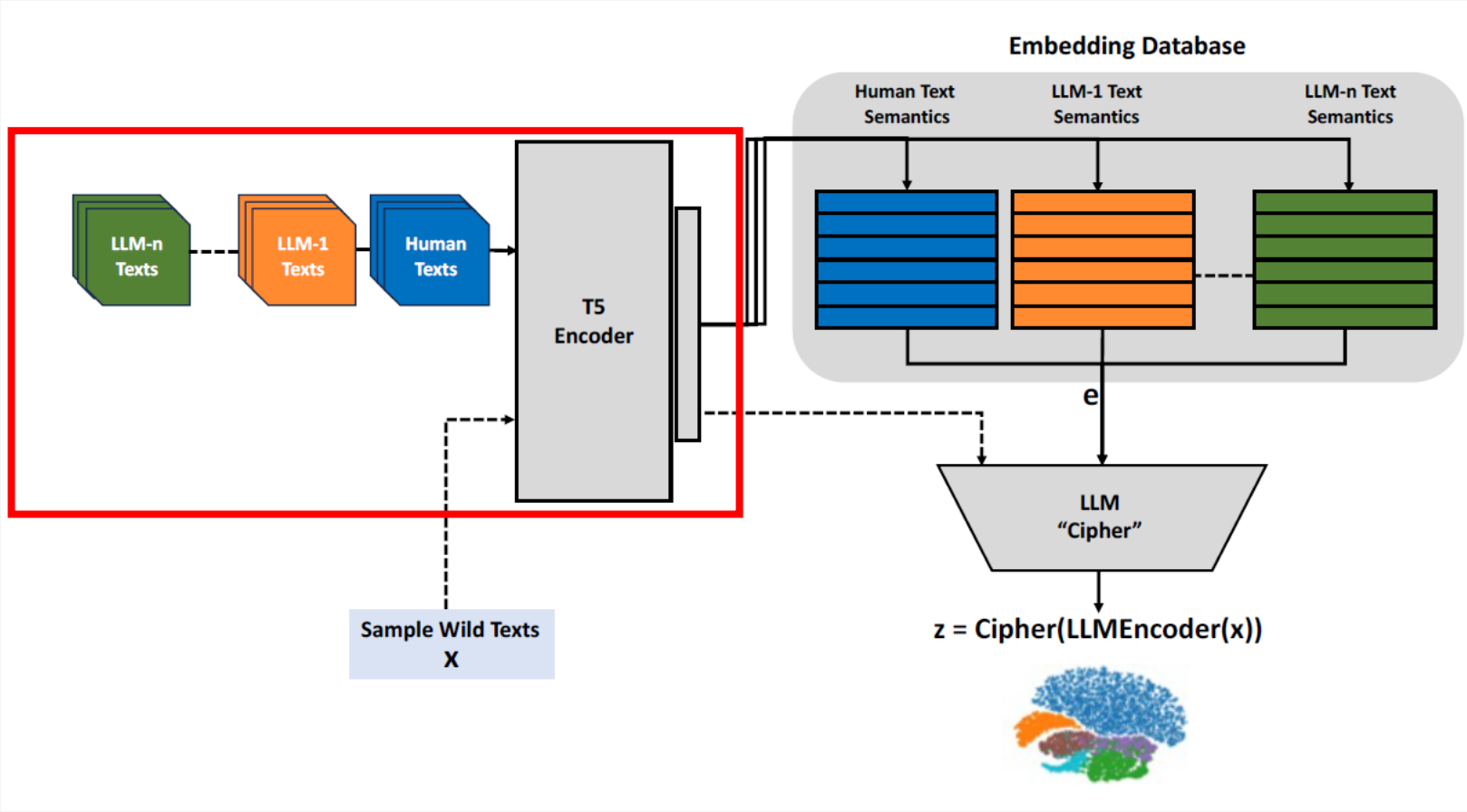
Our Approach - T5LLMCipher

- Leverages LLM encoders for rich text embeddings
- Embeddings are stored in database
- Identifies unique "fingerprints" of different text generators
- Goal: Robust detection across generators and domains



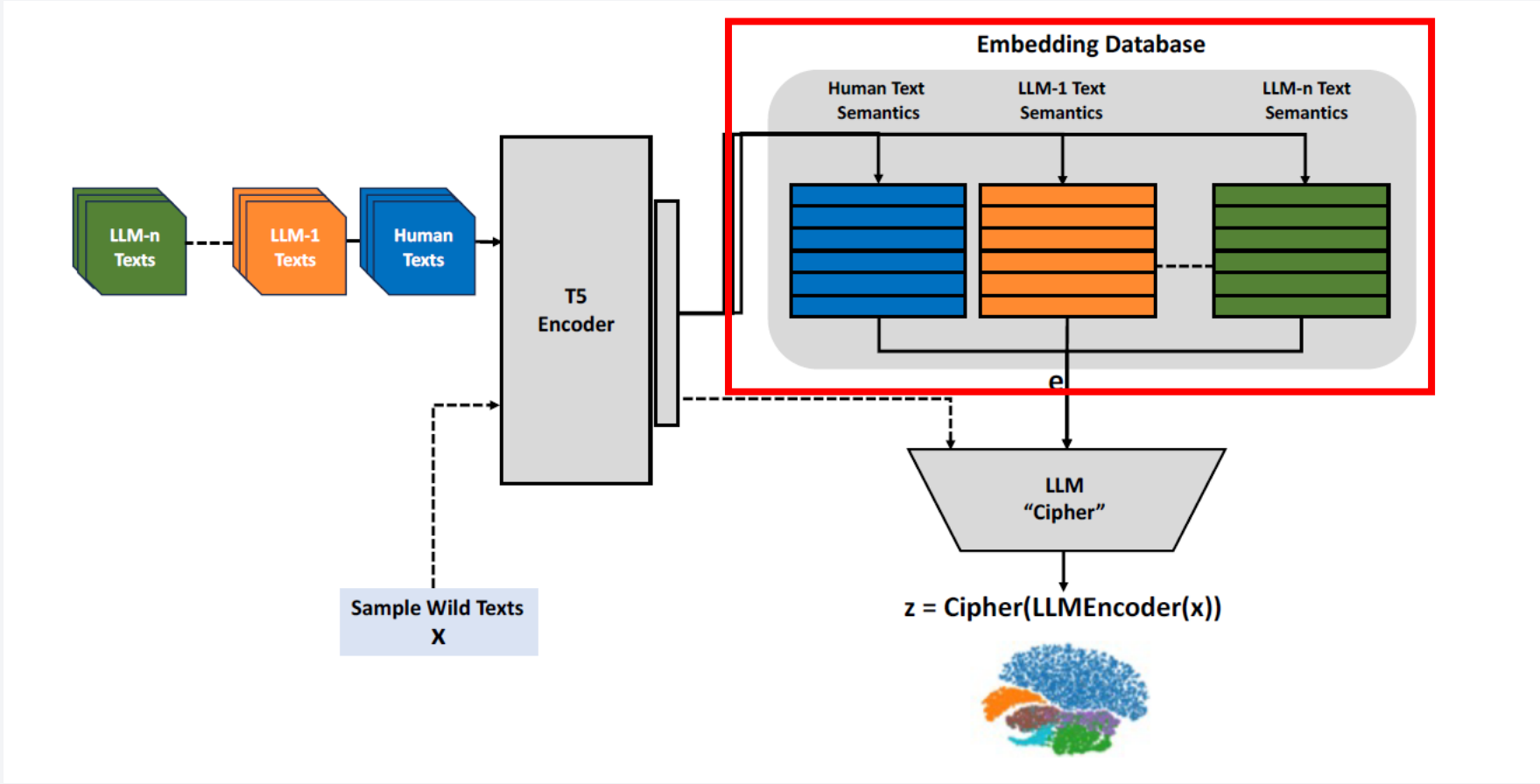
Our Approach - T5LLMCipher

- LLM Encoder: Transforms raw text into dense embeddings



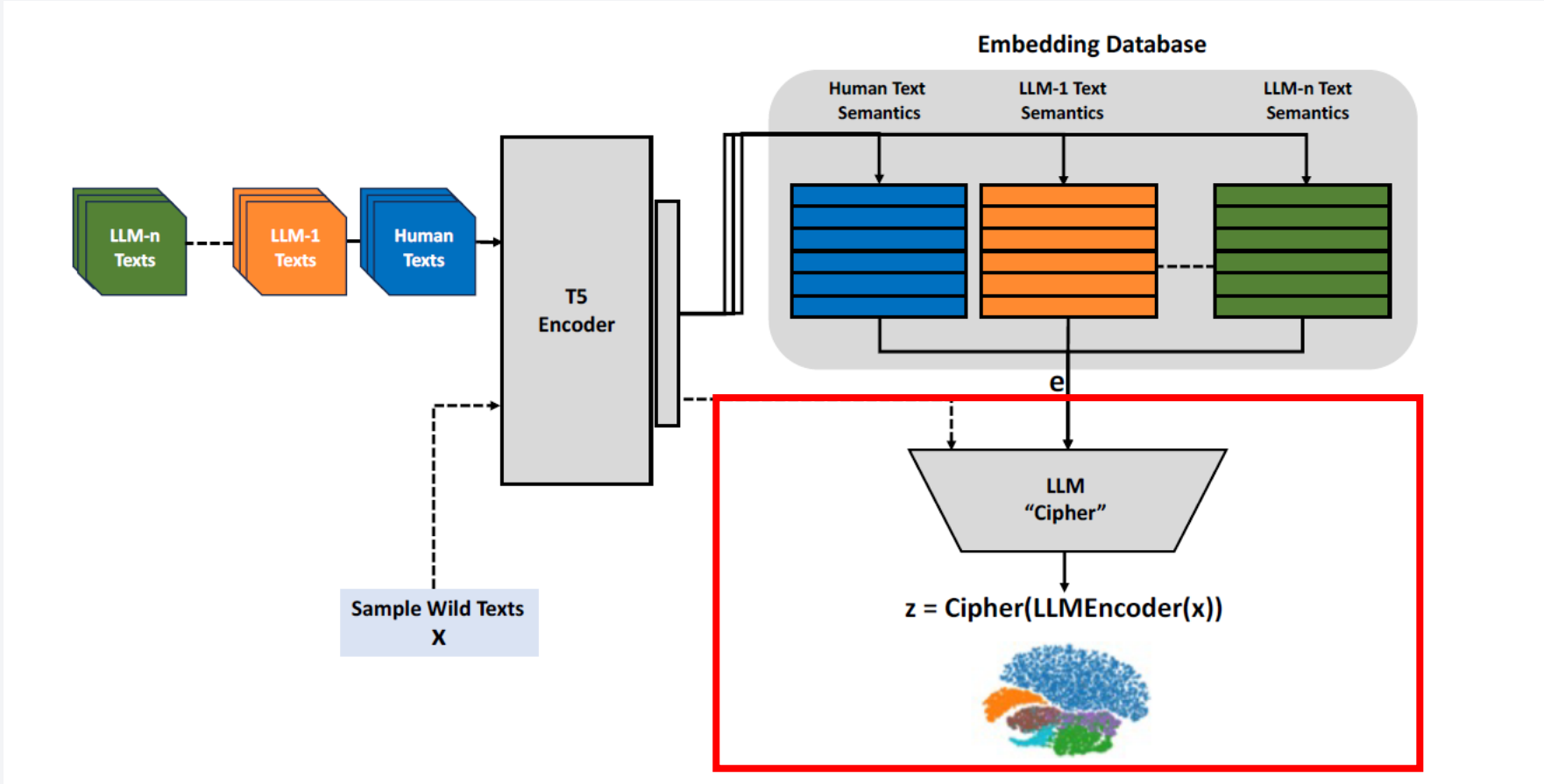
Our Approach - T5LLMCipher

- Embedding Database: Stores embeddings for efficient retrieval

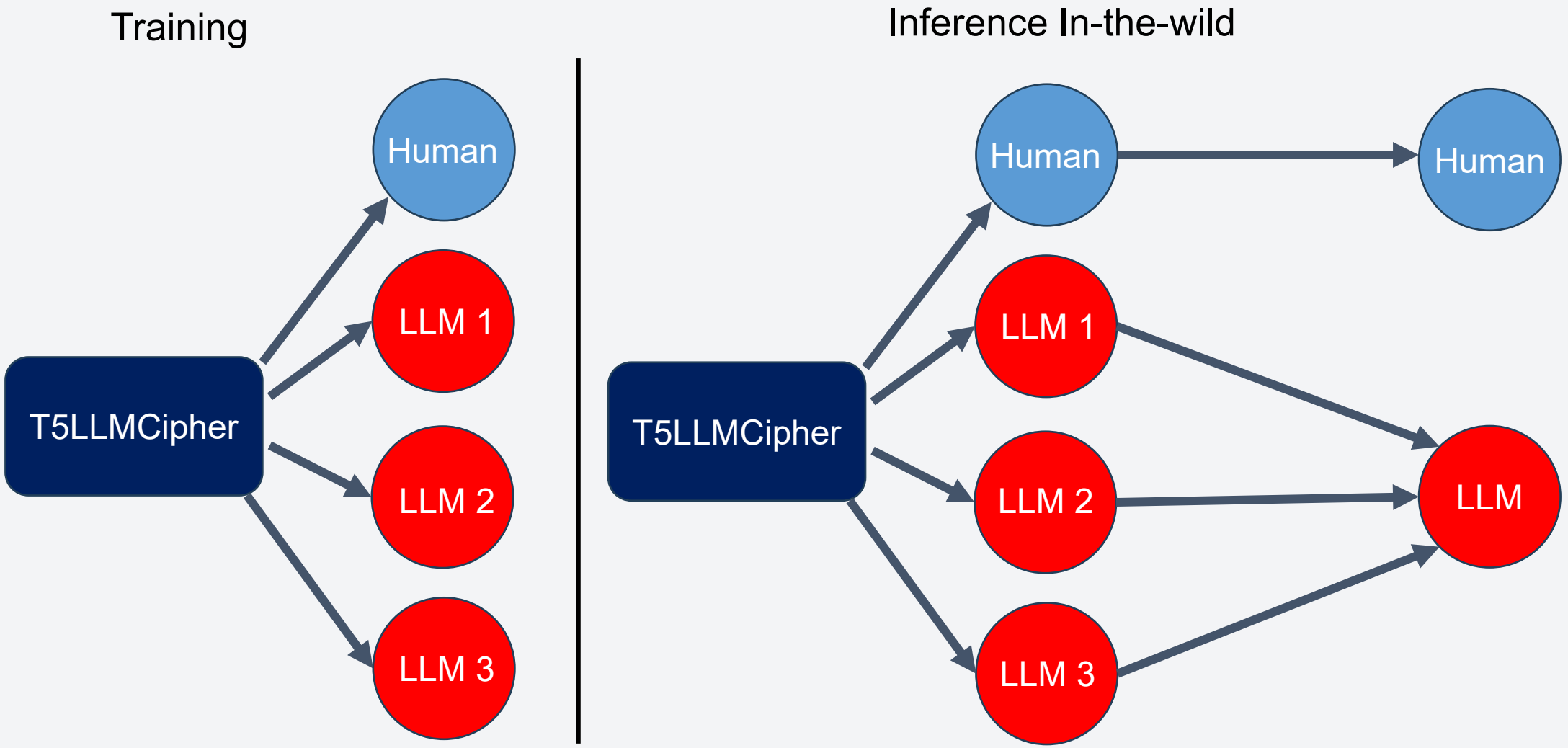


Our Approach - T5LLMCipher

- Classifier ("Cipher"): Maps embeddings to classification decisions
- Variants: MLP, KNN, Contrastive KNN



T5LLMCipher-MC



In-the-wild Experiment

- Models evaluated on **domains** and **generators** that they were **not trained** on
- Model trained for the task of **identifying source generator** outperforms those trained for **binary task** of human vs machine text

	AI Writer	Article Forge	Kafkai	Reddit Bot	Avg
T5LLMCipher-MC	62.3	84.4	61.8	84.6	**73.3**
T5LLMCipher-Bi	59.2	78.8	46.2	82.1	66.6
T5LLMCipher-C-KNN	57.2	83.2	49.2	82.9	68.1
T5LLMCipher-KNN	37.2	45.4	32.8	67.0	45.6
RoBERTaLLMCipher-MC	61.5	81.2	41.9	81.8	66.6
BERT-D	68.6	70.6	55.6	67.3	65.5
GLTR-BERT	49.2	83.8	29.0	73.8	59.0
GLTR-GPT2	29.6	83.1	29.2	83.5	56.4
RoBERTa-D	20.9	48.7	37.6	76.8	46.0
RADAR	50.2	61.6	60.2	18.0	48.0
Fast-DetectGPT	32.1	52.9	98.7	97.5	70.3

Adversarial Robustness Experiment

- Additionally attacked the in-the-wild data
- Model trained for the task of **identifying source generator** outperforms those trained for **binary task** of human vs machine text

	AI Writer	Article Forge	Kafkai	Reddit Bot	Avg
T5LLMCipher-MC	54.9	77.3	54.9	81.5	**67.2**
T5LLMCipher-Bi	62.4	66.9	22.8	50.0	50.5
T5LLMCipher-C-KNN	31.0	53.0	22.4	53.1	39.9
T5LLMCipher-KNN	19.8	23.7	37.3	37.3	29.5
BERT-D	54.6	55.2	52.1	58.8	55.2
GLTR-BERT	29.7	65.9	10.9	73.0	44.9
GLTR-GPT2	9.1	57.3	10.8	76.9	38.5
RoBERTa-D	13.0	36.6	29.1	82.6	40.3
RADAR	49.6	64.3	58.9	36.2	52.3
Fast-DetectGPT	29.1	54.0	54.0	97.7	58.7

Conclusion and Future Work

- **LLM Embeddings** can be used to distinguish between human and machine-generated texts
  - These models are more robust to unknown generators, domains, and adversarial attacks
- Models trained for the multi-class classification task of generator attribution can outperform models trained for binary task of human vs machine text
- Future Work
  - Adapt for multilingual machine text detection
  - More sophisticated “Cipher” architectures

Mazal Bethany,
PhD Candidate

mazal.bethany@utsa.edu

UTSA,
San Antonio, TX,
United States