



DEMASQ: Unmasking the ChatGPT Wordsmith

Kavita Kumari, Alessandro Pegoraro,
Hossein Fereidooni
Ahmad-Reza Sadeghi,
NDSS 2024

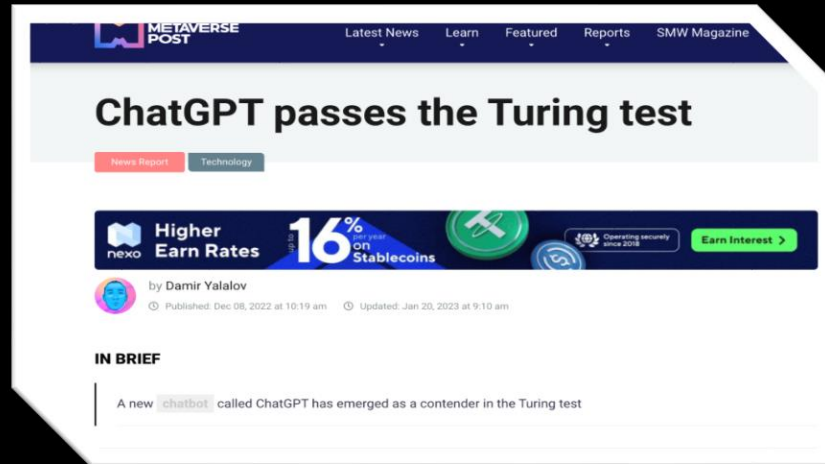


TECHNISCHE
UNIVERSITÄT
DARMSTADT



The AI Pandemic

The LL-Mania



METVERSE POST Latest News Learn Featured Reports SMW Magazine

ChatGPT passes the Turing test

News Report Technology

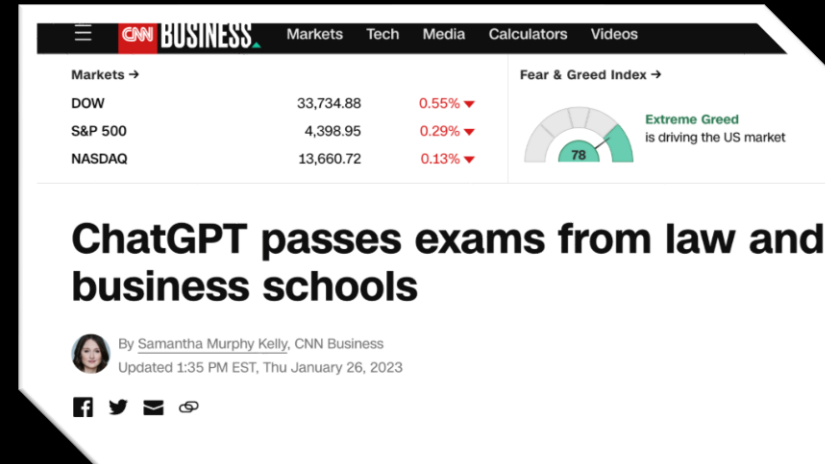
Higher Earn Rates 16% per year on Stablecoins [Earn Interest >](#)

Operating securely since 2018

by **Damir Yalalov**
Published: Dec 08, 2022 at 10:19 am Updated: Jan 20, 2023 at 9:10 am

IN BRIEF

A new `chatbot` called ChatGPT has emerged as a contender in the Turing test



CNN BUSINESS Markets Tech Media Calculators Videos

Markets →

DOW	33,734.88	0.55% ▼
S&P 500	4,398.95	0.29% ▼
NASDAQ	13,660.72	0.13% ▼

Fear & Greed Index →

78 Extreme Greed is driving the US market

ChatGPT passes exams from law and business schools

By **Samantha Murphy Kelly**, CNN Business
Updated 1:35 PM EST, Thu January 26, 2023

[f](#) [t](#) [e](#) [s](#)



MailOnline News

Home News Royals U.S. Sport TV&Showbiz Femall Health Science Money Travel Best Buys Discounts

Breaking News Australia Video Russia-Ukraine China Debate Meghan Markle Prince Harry King Charles III Weather Most read [Login](#)

The rise of the machines? ChatGPT CAN pass US Medical Licensing Exam and the Bar, experts warn - after the AI chatbot received B grade on Wharton MBA paper



RSNA
Radiological Society of North America

Membership Annual Meeting Journals Education

Search News

< RSNA News

ChatGPT Passes Radiology Board Exam

Study highlights growing potential of large language models in radiology

May 16, 2023

ChatGPT-Phobia

ARTIFICIAL INTELLIGENCE / TECH / POLICY

Top AI conference bans use of ChatGPT and AI language tools to write academic papers



/ AI tools can be used to 'edit' and 'polish' authors' work, say the conference organizers, but text 'produced entirely' by AI is not allowed. This raises the question: where do you draw the line between editing and writing?

By **James Vincent**, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Jan 5, 2023, 6:04 PM GMT+1 | [4 Comments](#) / [4 New](#)



Support the Guardian

Fund independent journalism with €5 per month

[Support us](#) →

The Guardian

[News](#) [Opinion](#) [Sport](#) [Culture](#) [Lifestyle](#) [More](#) ▾

[World](#) [UK](#) [Coronavirus](#) [Climate crisis](#) [Environment](#) [Science](#) [Global development](#) [Football](#) [Tech](#) [Business](#) [Obituaries](#)

Peer review and scientific publishing

● This article is more than 5 months old

Science journals ban listing of ChatGPT as co-author on papers

Some publishers also banning use of bot in preparation of submissions but others see its adoption as inevitable

Ian Sample Science editor



[NLP](#) [Chatbots](#) [Language models](#) [ML](#)

ChatGPT and AI Text Generator Tools Banned by ML Event

ICML authors cannot use text-generation tools unless 'presented as a part of the paper's experimental analysis.'



Ben Wodecki
January 6, 2023

🕒 2 Min Read



From ChatGPT to CheatGPT

REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Investigation

Disrupted

Top French university bans use of ChatGPT to prevent plagiarism

Reuters

January 27, 2023 7:21 PM GMT+1 · Updated 5 months ago

Bookmark Font Share

Tech

Japanese universities become latest to restrict use of ChatGPT

AI experts warn there could be disruptions in academia due to the breakthrough technology

Vishwam Sankaran • Monday 10 April 2023 09:08 • Comments



Id Scotland Health Education Technology Science Environment Business

Oxford and Cambridge ban ChatGPT over plagiarism fears but other universities choose to embrace AI bot

EXCLUSIVE

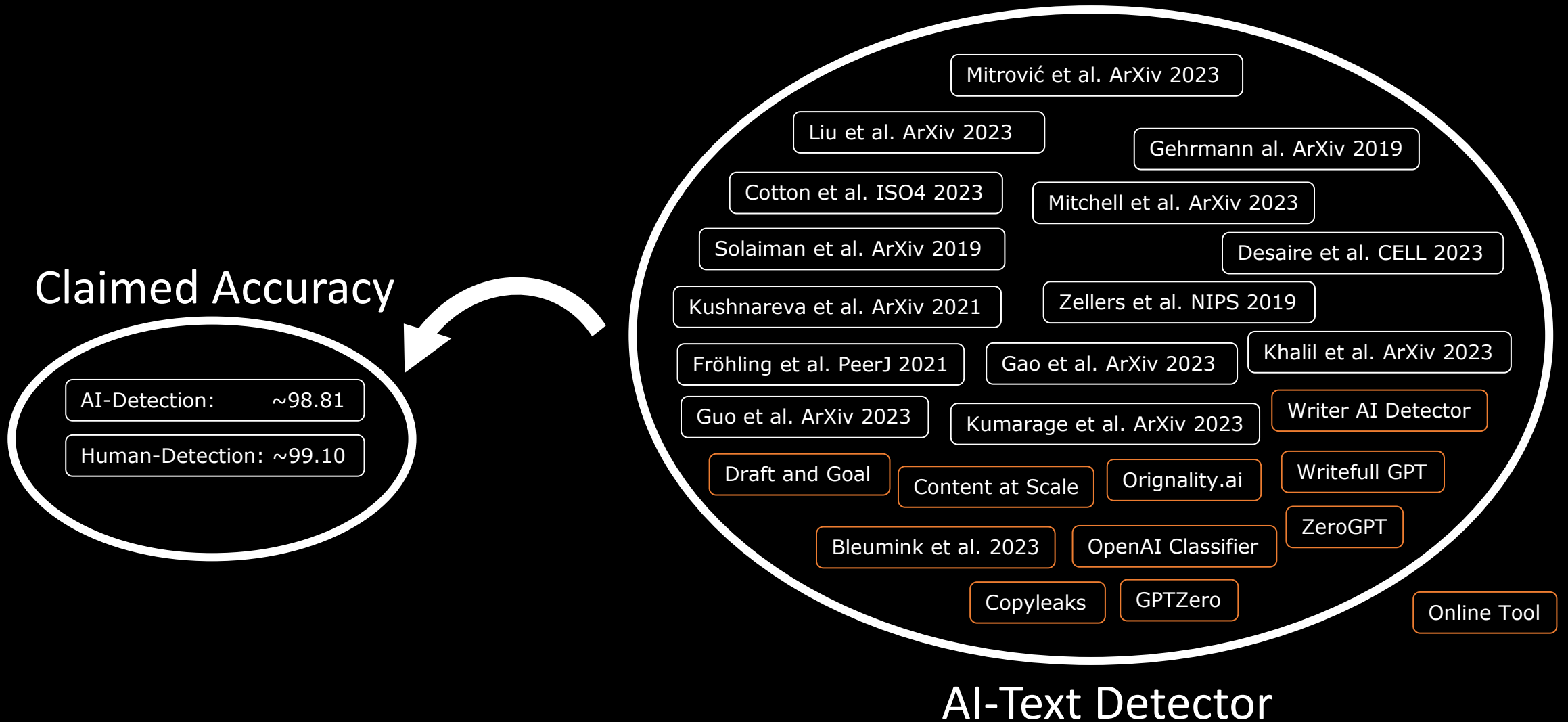
A third of Russell Group universities told i they have banned the AI chatbot for assessed essays, but other institutions have decided to explore its use for writing bibliographies and references





AI-Text Detection

AI-Text Detectors: State-of-the-Art

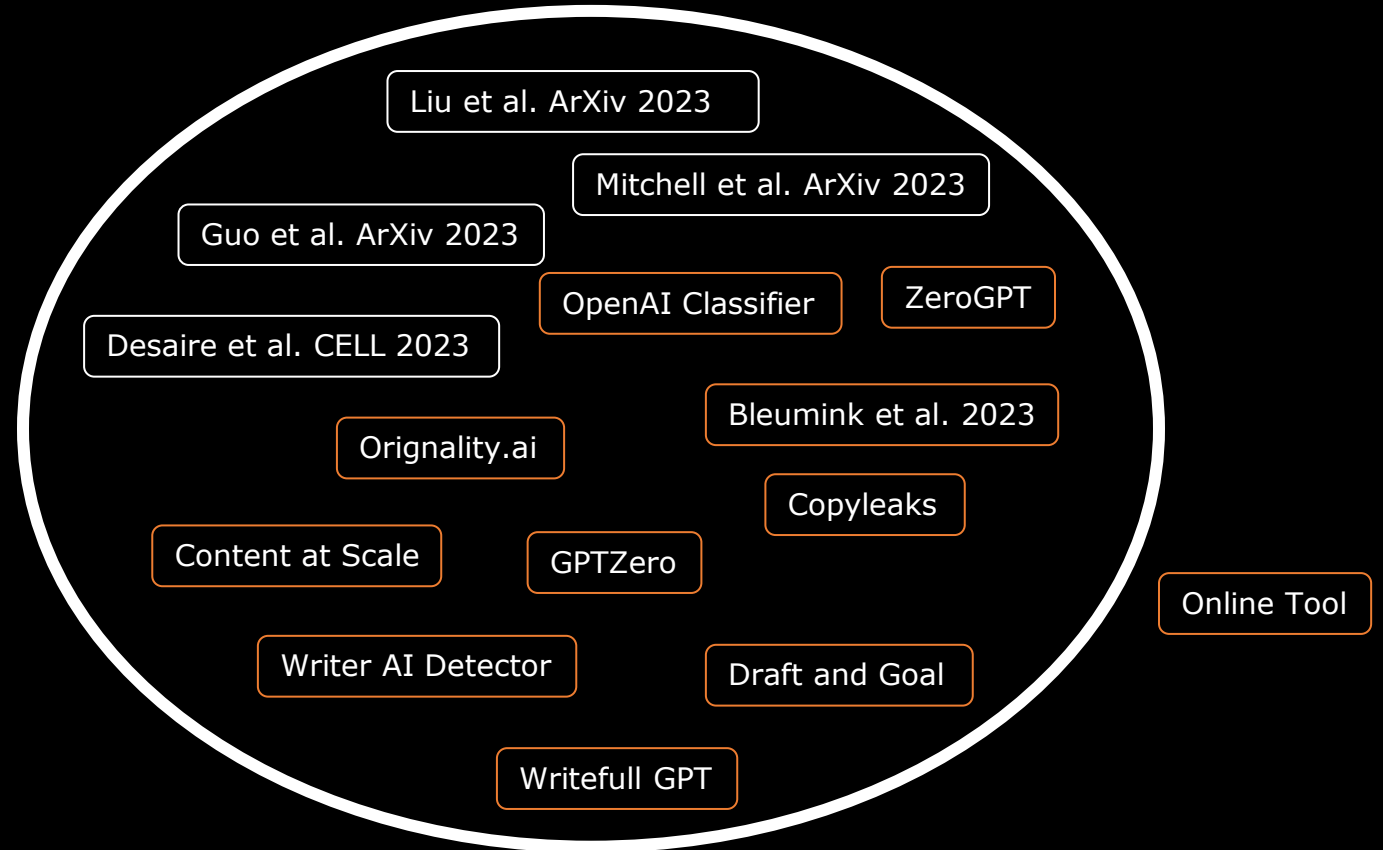


Existing ChatGPT Detectors

Detection Accuracy

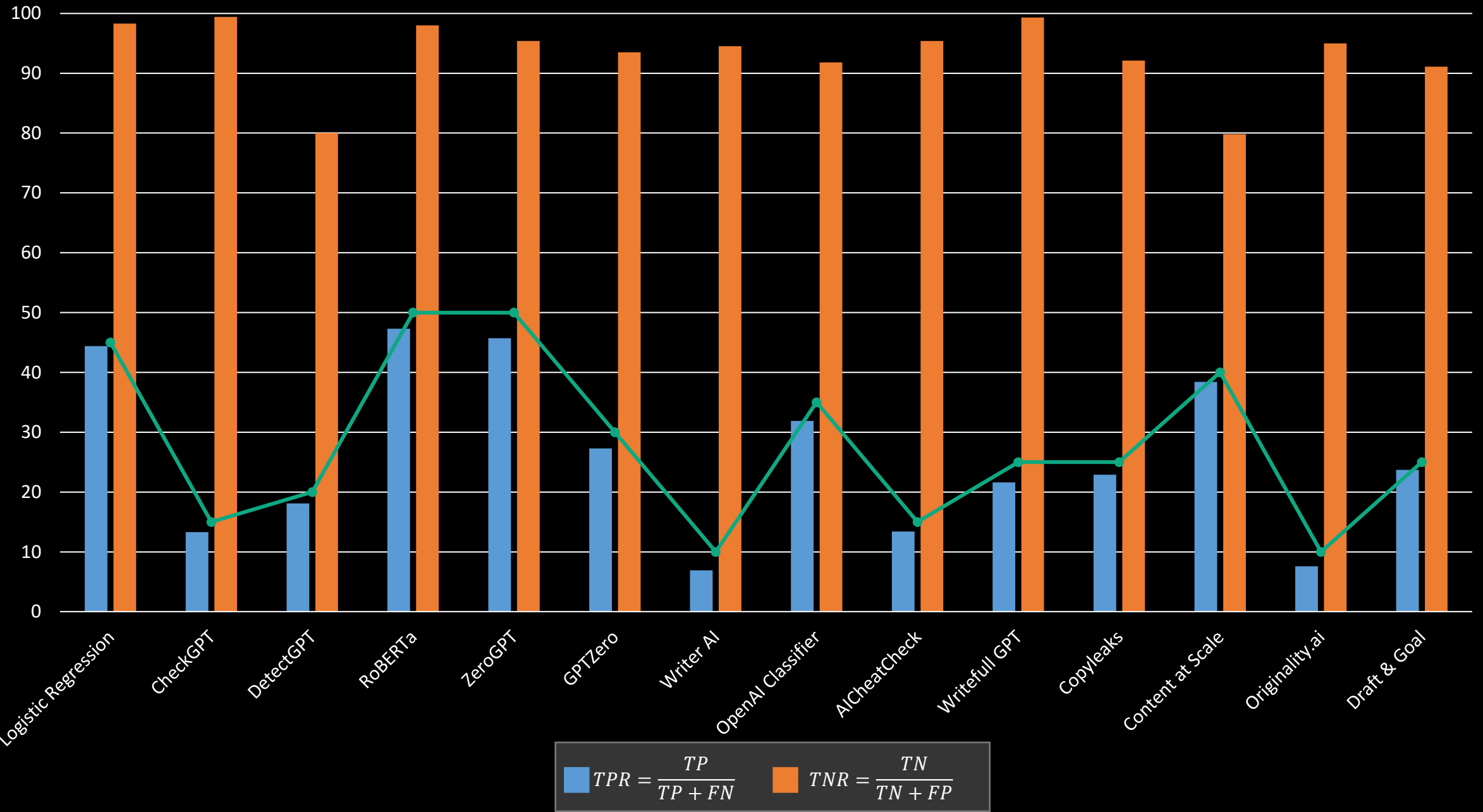
ChatGPT-Detection: ~97.87

Human-Detection: ~98.51



ChatGPT Detectors

Evaluation of ChatGPT Detectors



Human vs. Machine Behavior



Human Behavior Bias

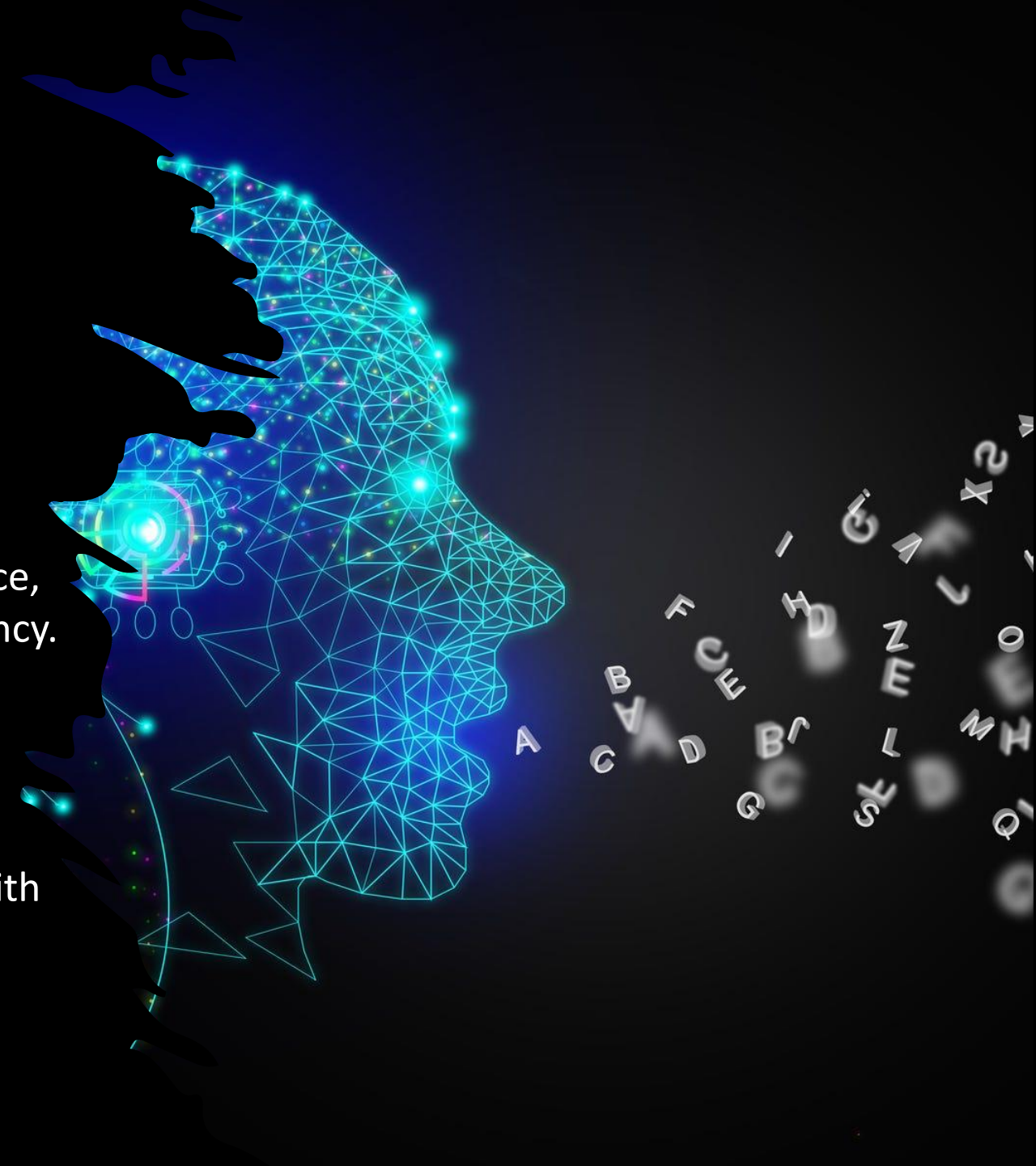
- Conscious
- Decision Making
- Creativity and Imagination
- Contextual Understanding
- Adaptability and Learning
- Cognitive differences
- Empathy and Emotional Intelligence
- Common Sense Reasoning
- Intuition and Instinct
- Humor and Sarcasm
- Psychology

Actions, Feelings, and Emotions



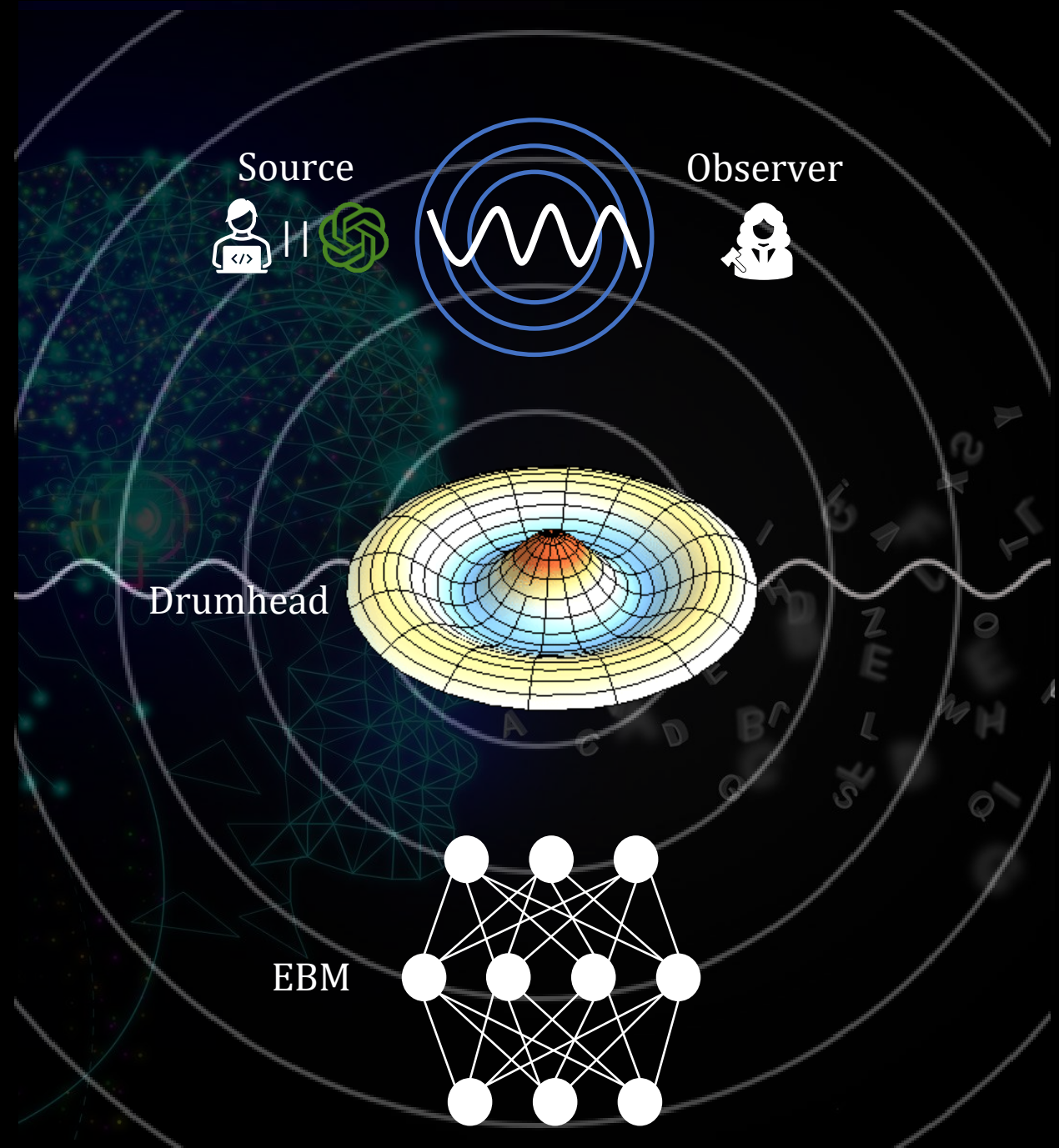
Intuition and Hypothesis

- Human interactions exhibit diverse patterns and inherent biases. Thus, when thoughts (or when translated into text) compared to phonetic waves vibrates with different frequency.
- Machines interact based on predefined rules. Hence, in this case, text strings vibrates with same frequency.
- Higher frequency \propto Higher energy and vice versa.
- Thus, human produced content vibrates with high energy and machine generated content vibrates with low energy.



Modeling the Basic Idea

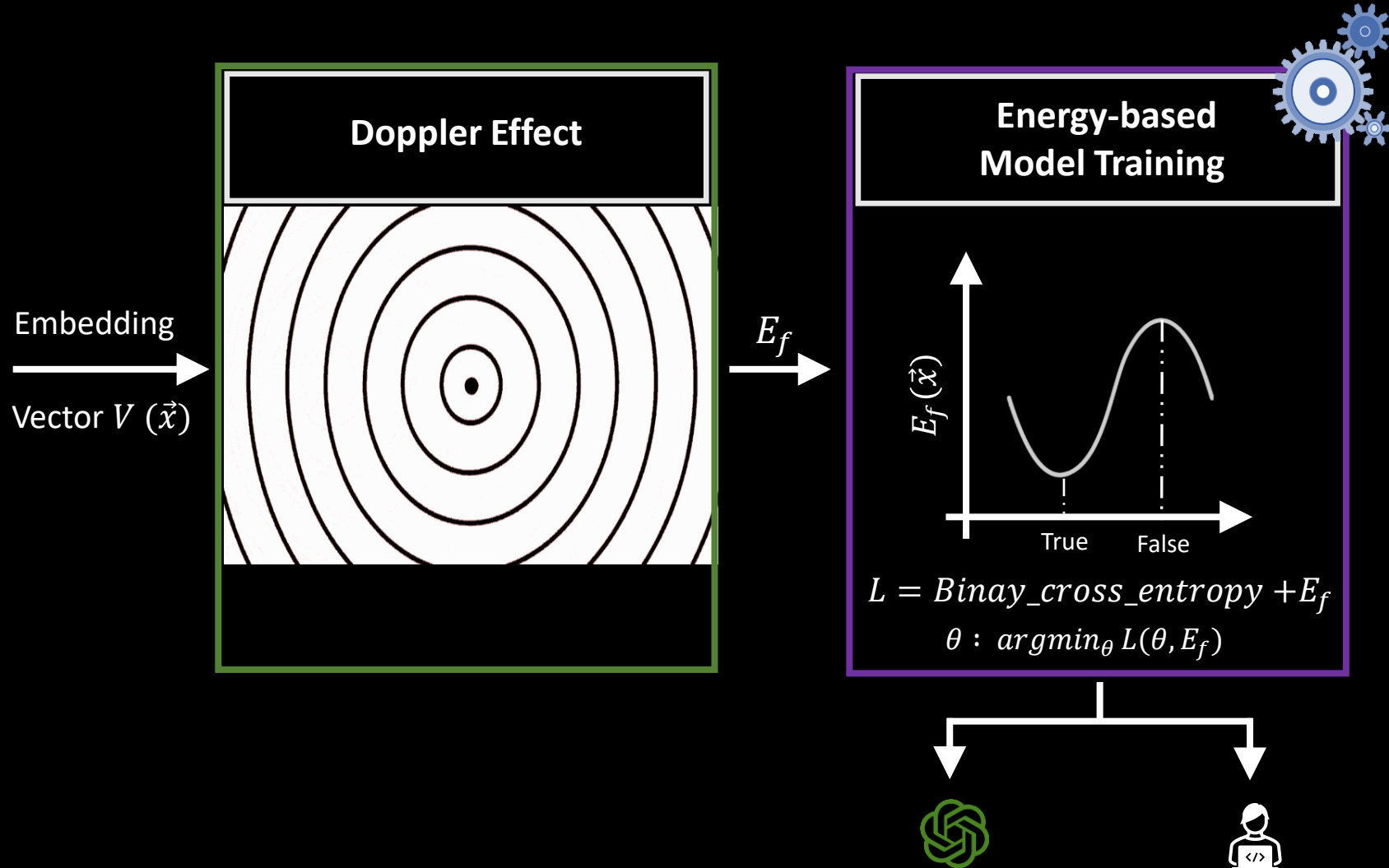
- Doppler effect involves source waves radiating outward in concentric circles towards the Observer.
- Model source frequencies as drumhead vibrations. Diametric waves propagate in concentric circles towards a fixed outer boundary.
- Adopt drumhead vibrations to model waves and source frequency.
- Train an Energy-Based Model (EBM) and integrate the observed frequency as an energy value.
- Finally, using XAI to counter hybrid rephrasing.





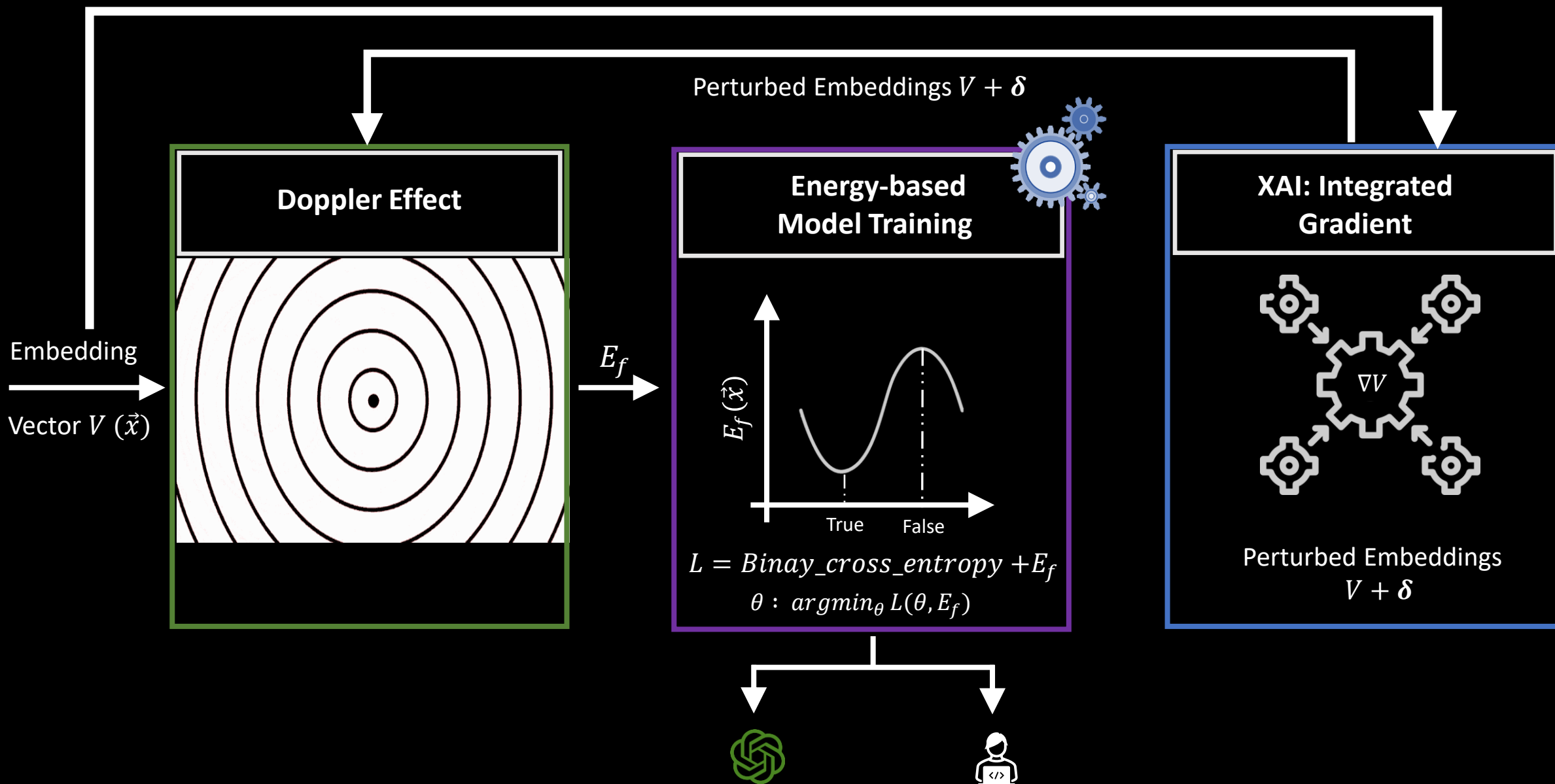
DEMASQ: Details

Core Components

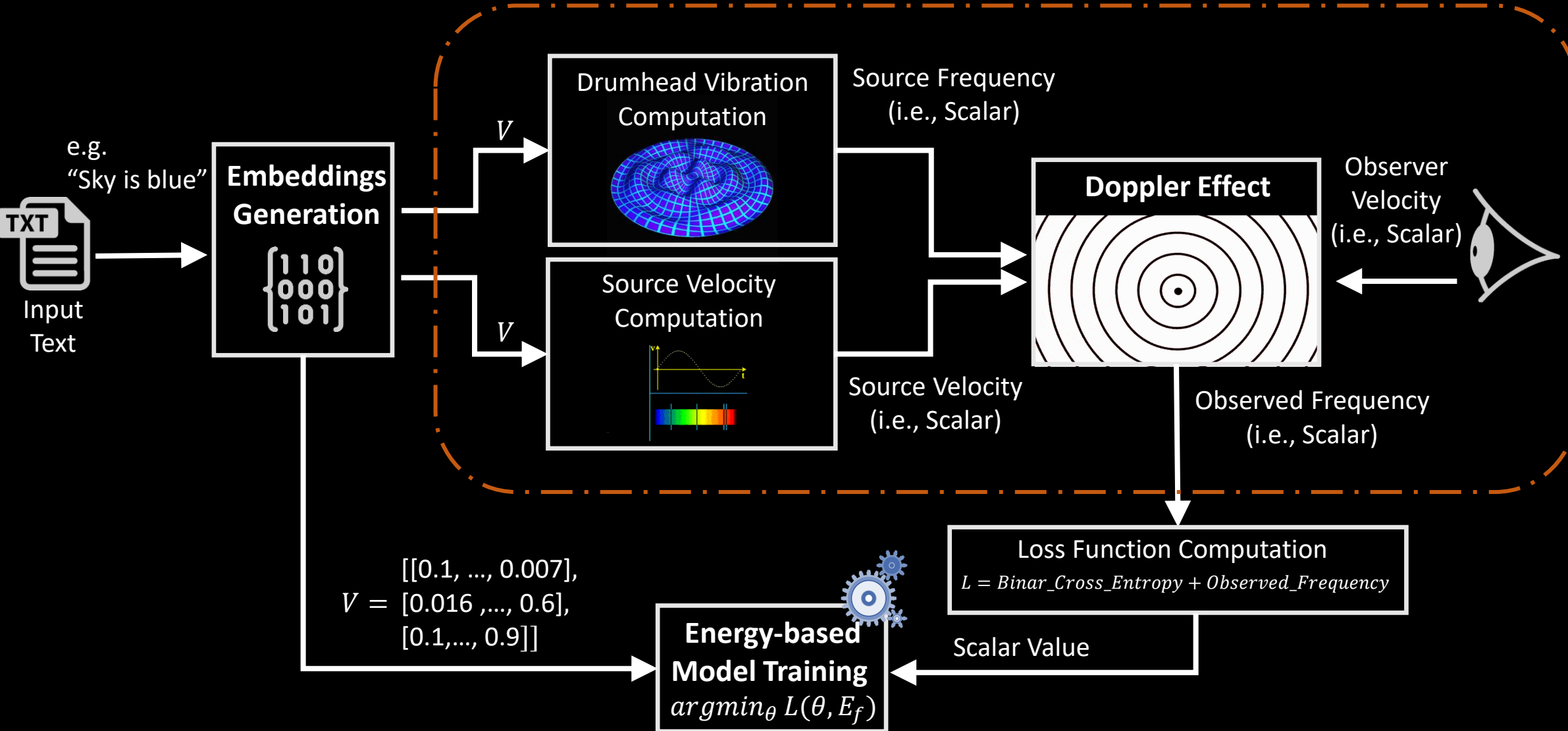


Core Components

Embedding $V := [v_1, v_2, \dots, v_k]$



DEMASQ: Training

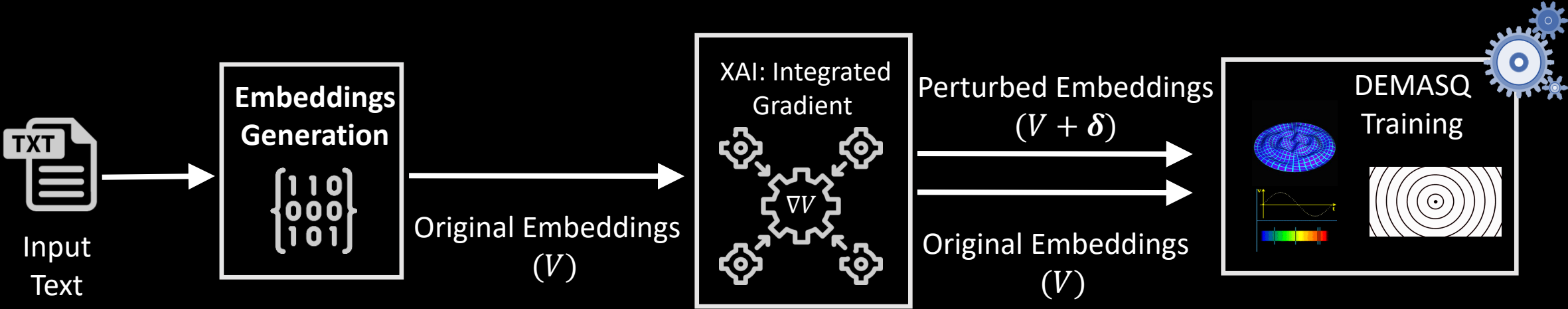


DEMASQ: Hybrid Text

Generate hybrid text
(Zombie text)

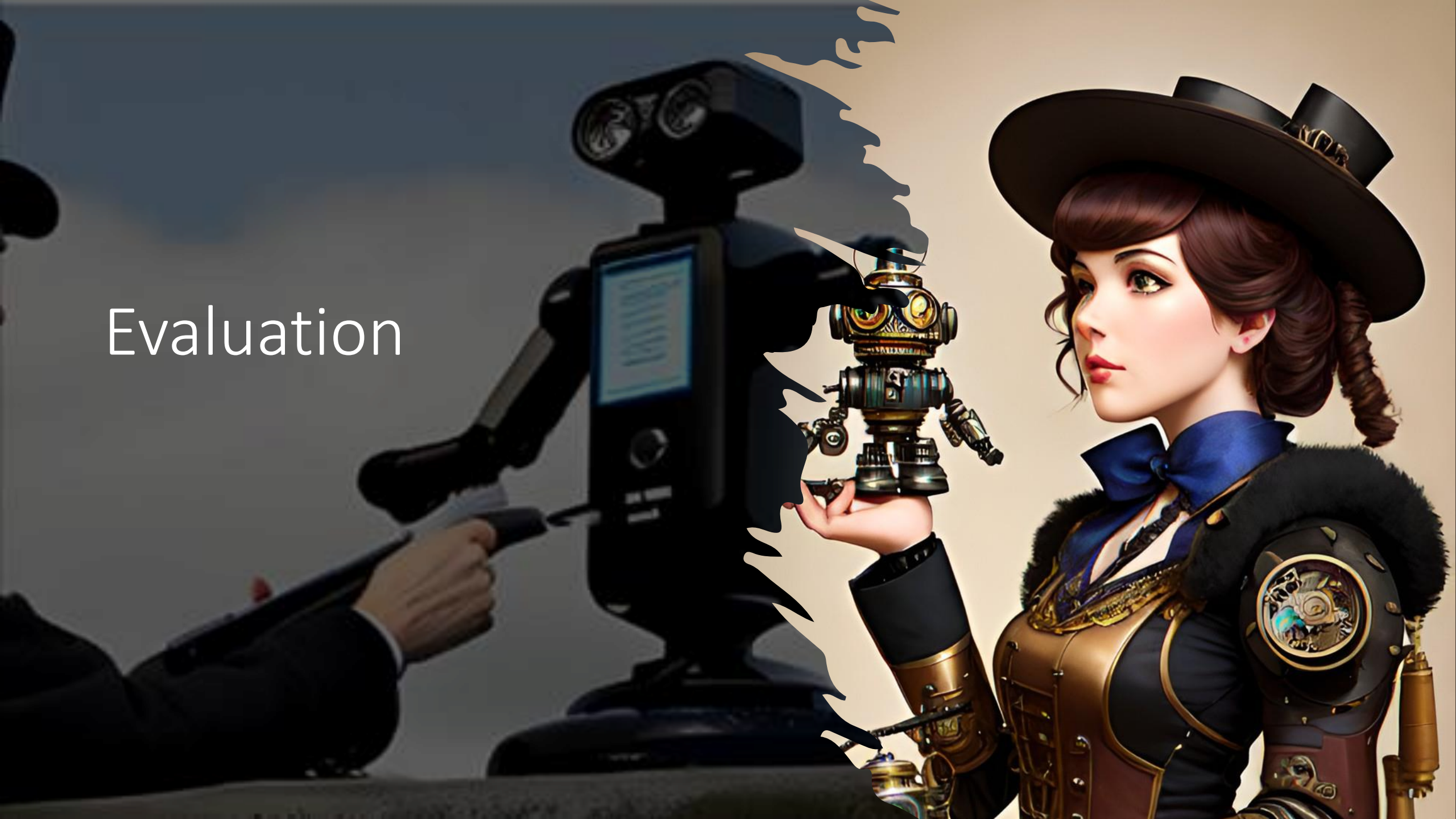


- Blue is the dominant color of the **sky**
- The sky's color **is** a vibrant shade of blue
- The sky's color exhibits a vibrant shade **of blue**
- The **sky's** color exhibits a vibrant shade of blue



e.g., “The color of the sky is vibrant blue”

Evaluation



Evaluation








$$\text{TPR} = \frac{TP}{TP + FN}$$

ChatGPT = 0

$$\text{TNR} = \frac{TN}{TN + FP}$$

Human = 1

Results for DEMASQ trained on different datasets


	Dataset	Records	TPR (%)	TNR (%)
	Medical	4,992	96.6	96.4
	Finance	15,732	94.6	92.7
	News	3,620	95.2	87.7
	Research	2,664	87.1	91.7
	Q&A	4,748	73.9	75.4
	Wiki	3,368	78.5	85.8
	Blog	99,054	100.0	100.0
	Total	134,178	97.0	96.5

Evaluation

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{TNR} = \frac{TN}{TN + FP}$$

ChatGPT = 0 Human = 1

Results for DEMASQ trained on different datasets of abstracts

 Liu et al. arXiv 2023	Records	TPR (%)	TNR (%)
TASK 1: Create Abstract from Title	100,000	94.7	97.9
TASK 2: Complete Abstract from initial segment	100,000	88.7	88.4
TASK 3: Rephrase Abstract	100,000	84.3	81.0



BIAS in TASK 1

Record	Label
99990	1
99991	0
99992	1
99993	0
99994	1
99995	0
99996	1
99997	0
99998	1
99999	0



BIAS in TASK 3

Record	Label
99990	1
99991	0
99992	1
99993	0
99994	1
99995	0
99996	1
99997	0
99998	1
99999	0

Evaluation

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{TNR} = \frac{TN}{TN + FP}$$

Results for DEMASQ and CheckGPT on the rephrased datasets

Rephrased Liu et al. arXiv 2023	Records	CheckGPT		DEMASQ	
		TPR (%)	TNR (%)	TPR (%)	TNR (%)
TASK 1: Create Abstract from Title	100,000	82.7	99.7	93.0	94.1
TASK 2: Complete Abstract from initial segment	100,000	56.3	96.1	83.4	84.5
TASK 3: Rephrase Abstract	100,000	3.8	99.0	74.9	75.1

Conclusion

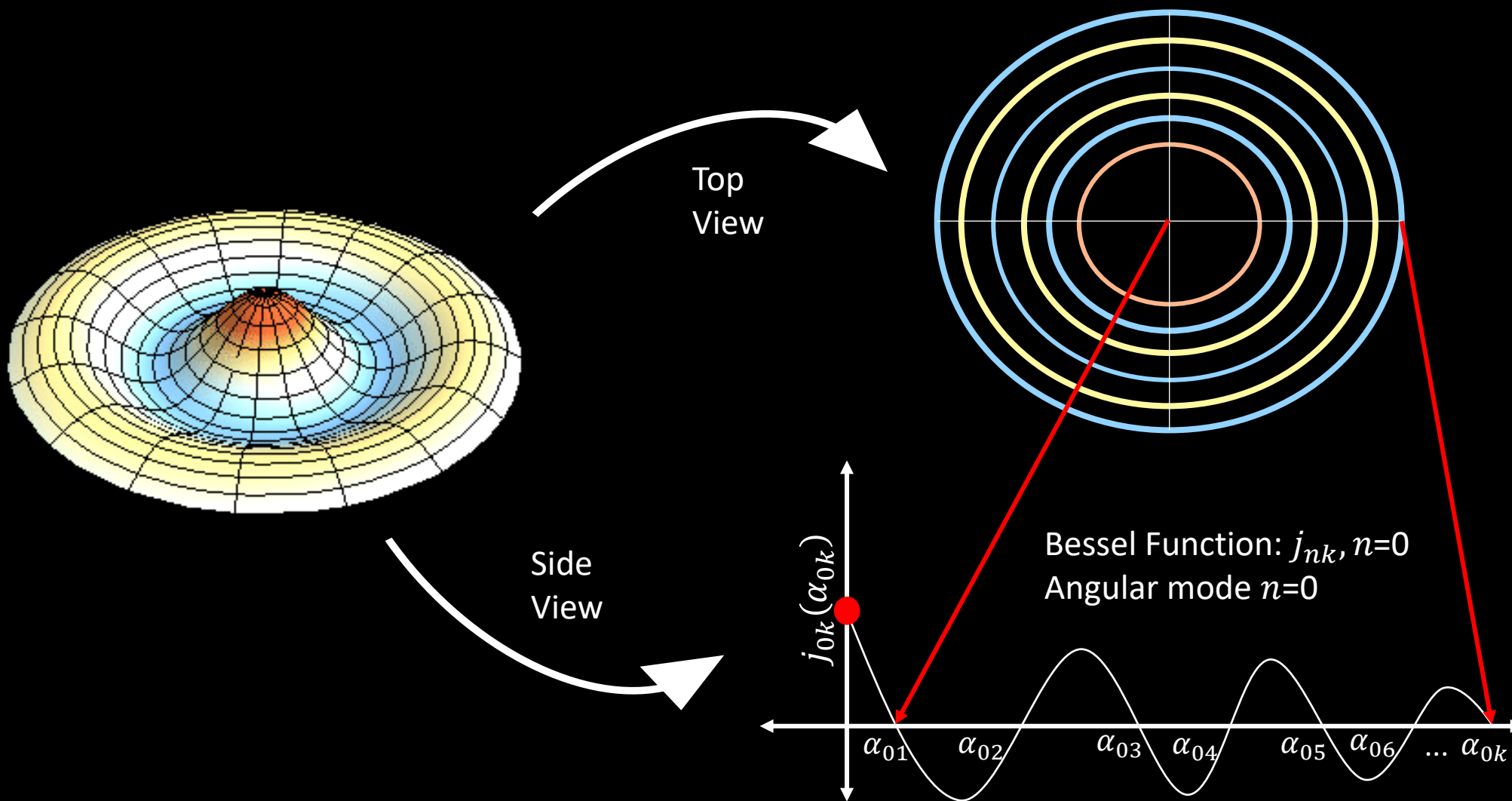
- Proposed a novel detection method DEMASQ to accurately detect the ChatGPT-generated text.
- Incorporates the effects introduced by various factors during human interactions with other humans or machines.
- Considers rephrasing techniques employed by humans to modify ChatGPT-generated responses and evade detection.
- Evaluations on a comprehensive benchmark dataset comprising over 100,000 samples, which covers diverse domains.



Additional
information

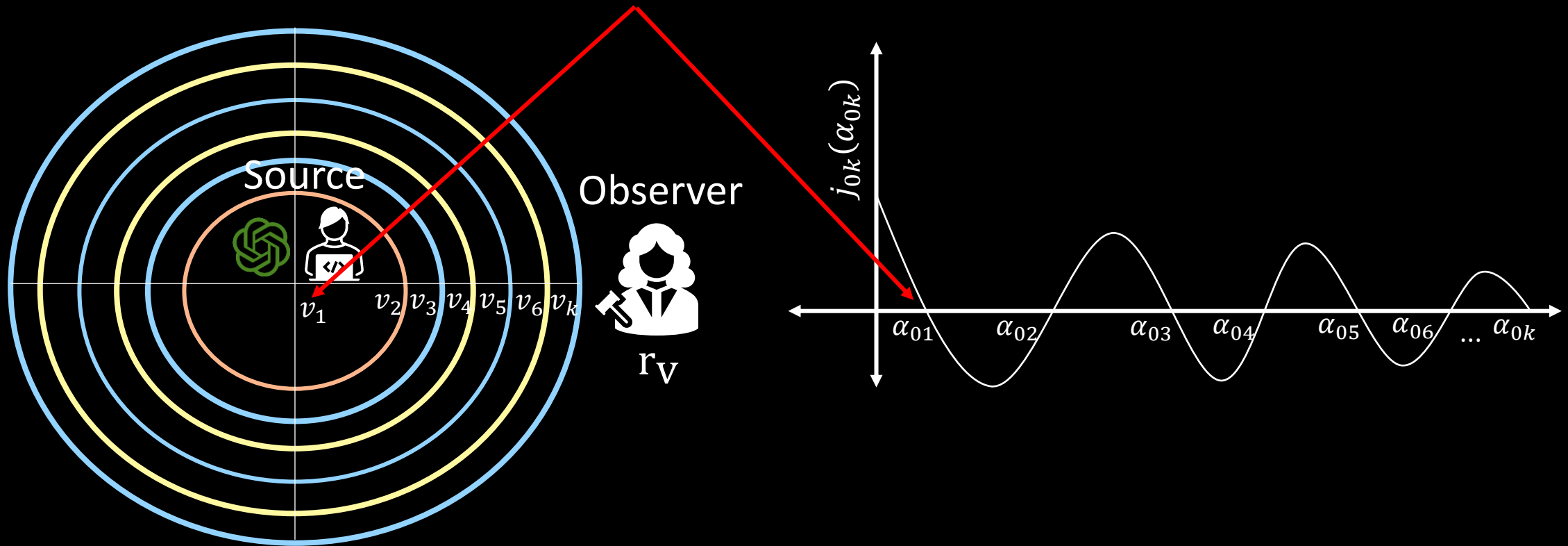


Drumhead Vibration



Analogy: Computing Observed Frequency

Embedding $V = [v_1, v_2, \dots, v_k]$



Source Frequency

$$E_{f_S} = \frac{\alpha_{0k}}{\alpha_{01}}$$



Observed Frequency

$$E_f = \frac{c_V + r_V}{c_V - s_V} * E_{f_S}$$

Doppler Effect

S_V
(Source Velocity (H))



$$\text{var}(V(\vec{x}))$$



S_V
(Source Velocity (C))



$$0$$



r_V
(Observer Velocity)



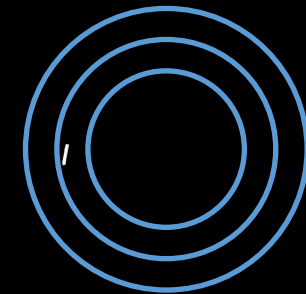
$$0.8$$



C_V
(Medium Velocity)



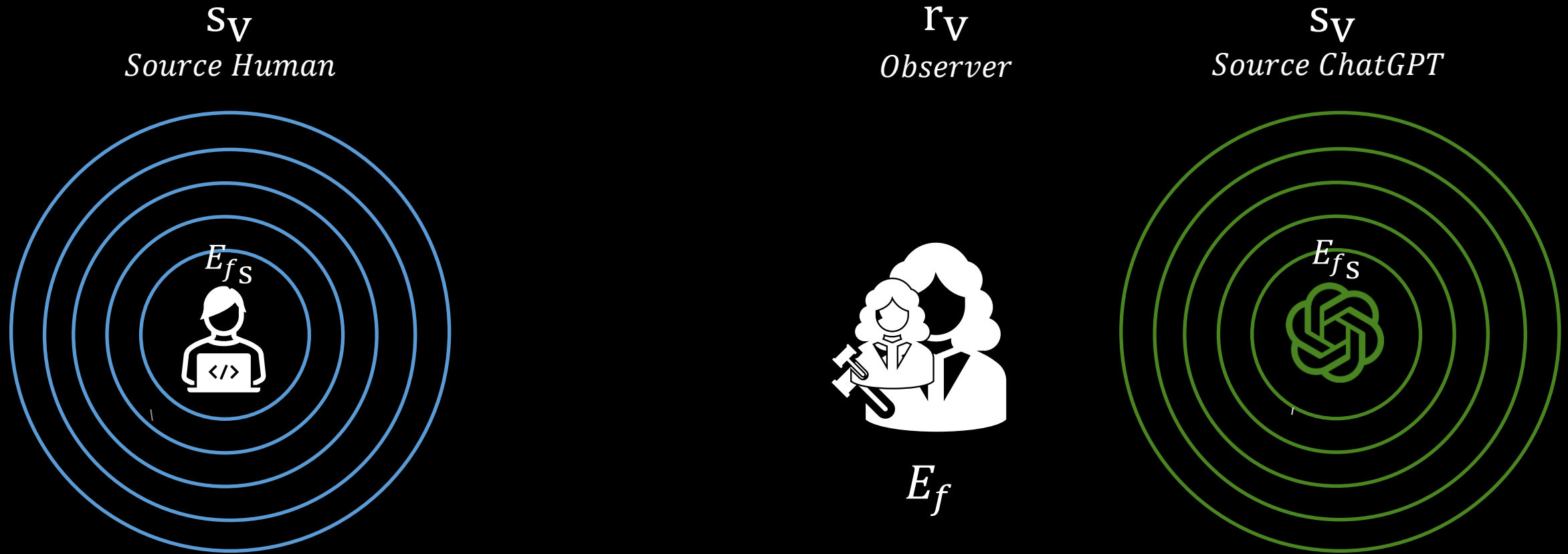
$$\text{var}(V(\vec{x})) * J_{0k}(\alpha_{01})$$



Observed Frequency

$$E_f = \frac{C_V + r_V}{C_V - S_V} * E_{f_S}$$

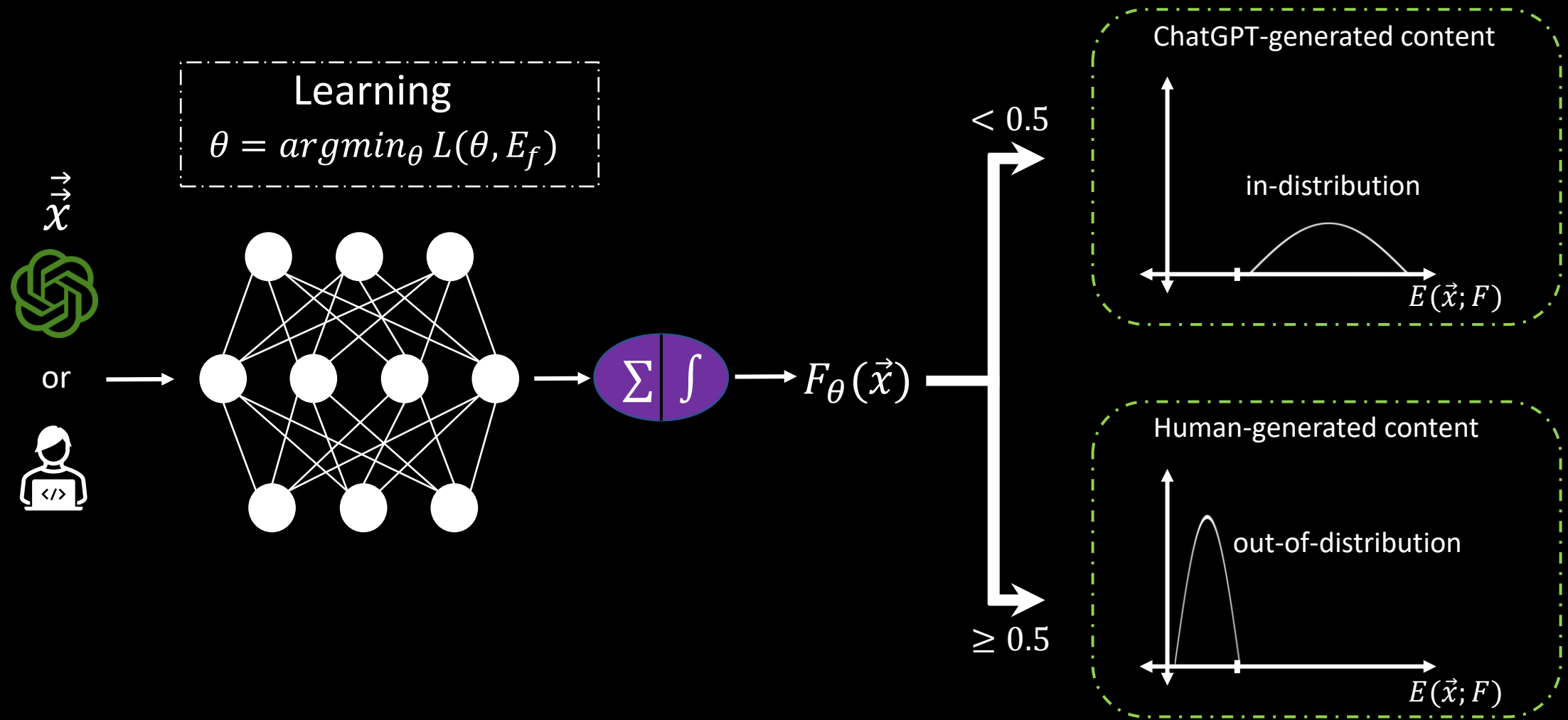
Doppler Effect



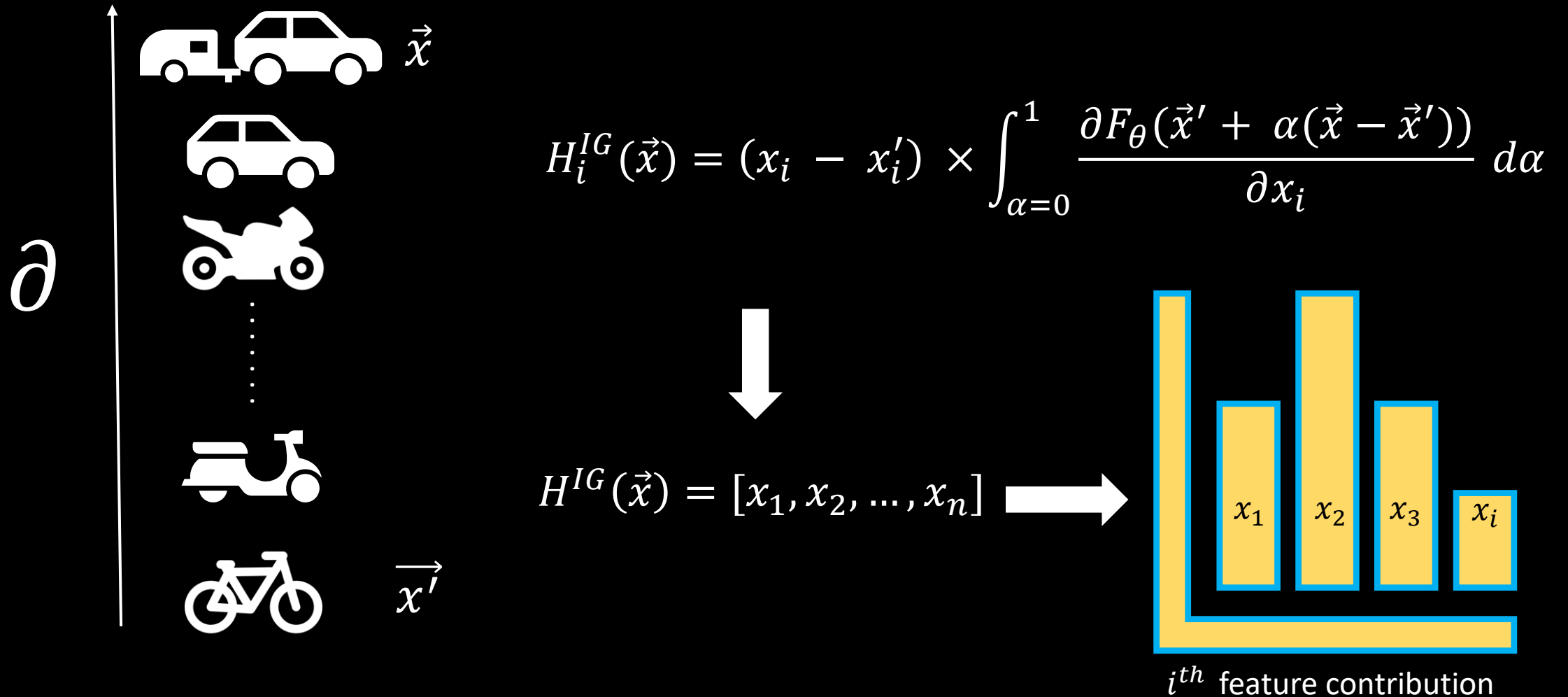
$$E_f = \frac{c_V + r_V}{c_V - S_V} * E_{fS} \longrightarrow$$

E_{fS} : *Source (ChatGPT or human) frequency*
 c_V : *Medium velocity*
 E_f : *Observed frequency*

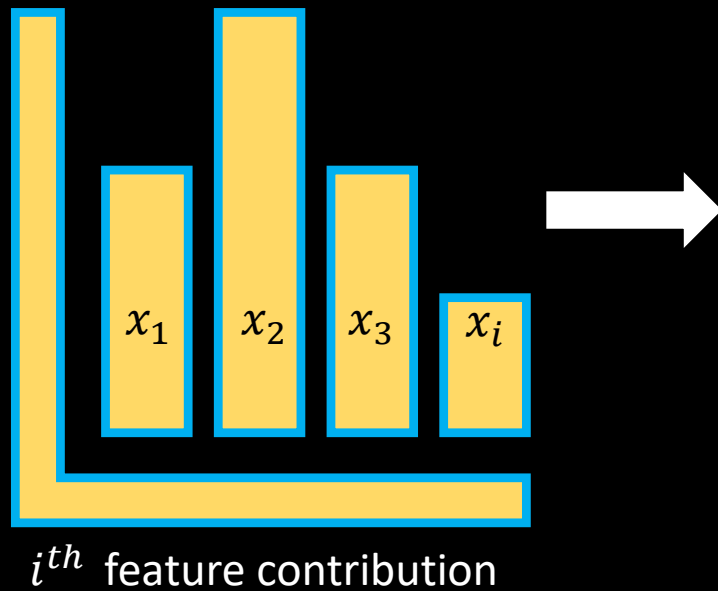
Energy-Based Model (EBM)



Integrated Gradient



Perturbing V



$$H^{IG} : [x_2, x_3, x_1, x_{71}, \dots, x_i]$$

$$V + \epsilon_1 : [v_1, v_2, v_3, v_4, \dots, v_i]$$

$$H^{IG} : [x_2, x_3, x_1, x_{71}, \dots, x_i]$$

$$V + \epsilon_2 : [v_1, v_2, v_3, v_4, \dots, v_i]$$

...

$$H^{IG} : [x_2, x_3, x_1, x_{71}, \dots, x_{20}, \dots, x_i]$$

$$V + \epsilon_{20} : [v_1, v_2, v_3, v_4, \dots, v_{20}, \dots, v_i]$$