SOPHOS

Multimodal AI: the sixth sense for cyber defence

Younghoo Lee
Principal Data Scientist @ Sophos

VB2024

Younghoo Lee

Principal Data Scientist @ Sophos

Enthusiast of hiking and running

Younghoo.lee@Sophos.com

https://www.linkedin.com/in/younghoolee

Agenda

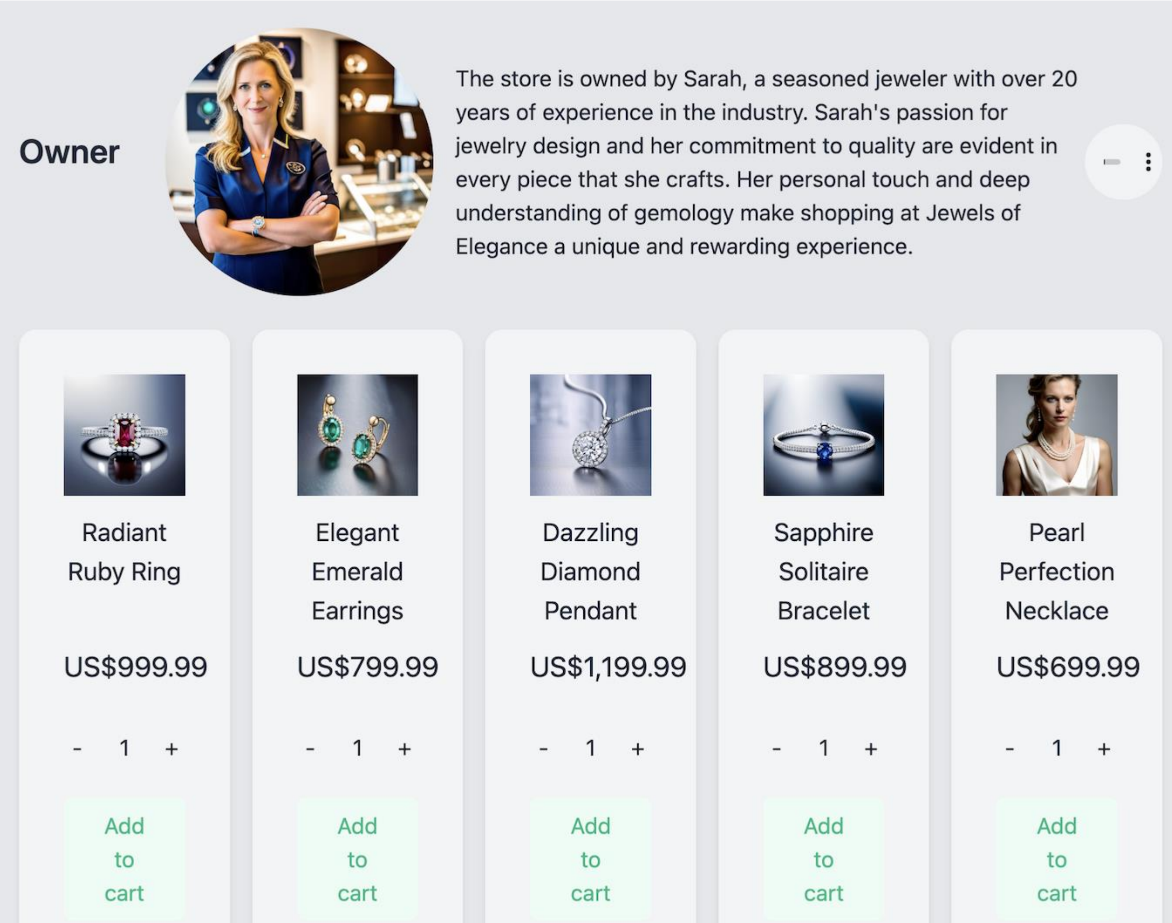
- Introduction to Multimodal AI
- Multimodal AI in Cybersecurity
  - Phishing email detection
  - NSFW website classification
- Conclusion

SOPHOS

Multimodal AI

VB2024

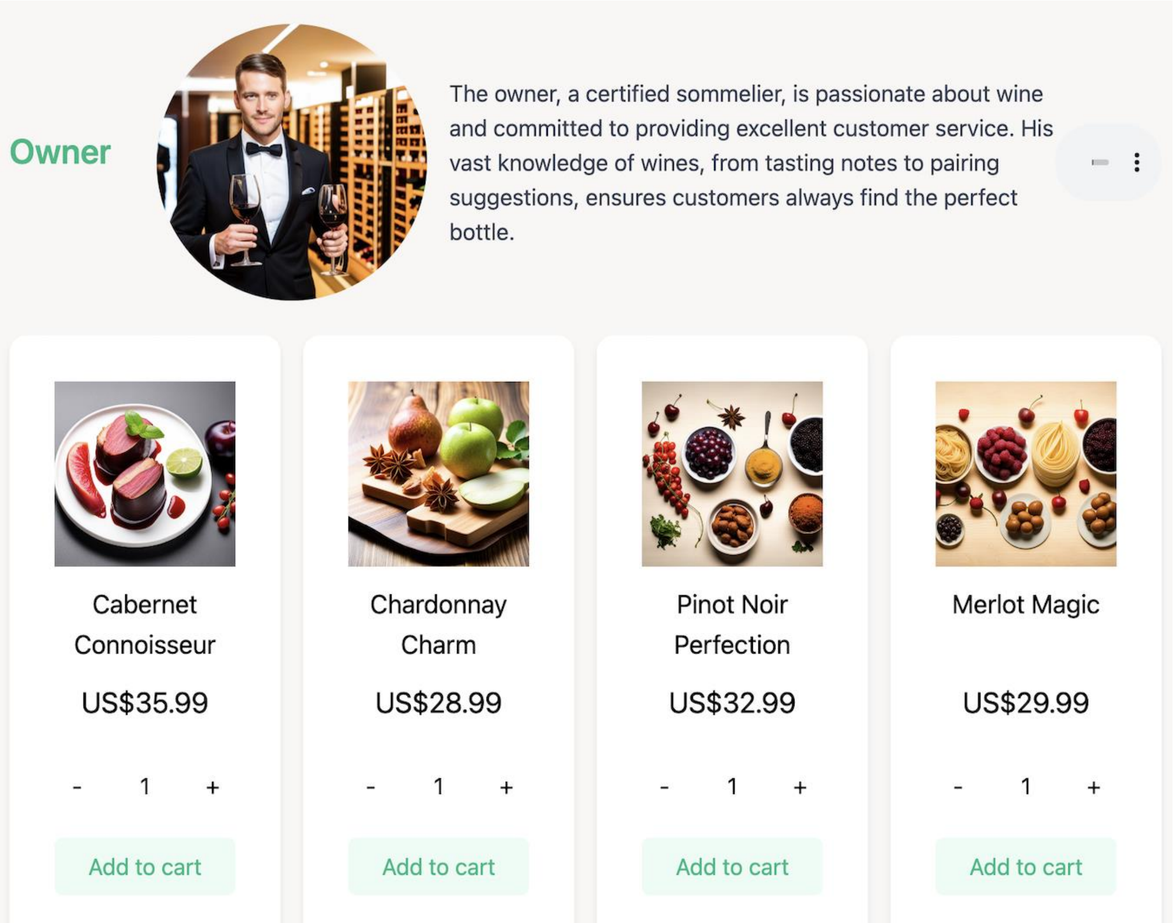
Cybersecurity Challenges: Malicious websites



Owner

The store is owned by Sarah, a seasoned jeweler with over 20 years of experience in the industry. Sarah's passion for jewelry design and her commitment to quality are evident in every piece that she crafts. Her personal touch and deep understanding of gemology make shopping at Jewels of Elegance a unique and rewarding experience.

Radiant Ruby Ring	Elegant Emerald Earrings	Dazzling Diamond Pendant	Sapphire Solitaire Bracelet	Pearl Perfection Necklace
US$999.99	US$799.99	US$1,199.99	US$899.99	US$699.99



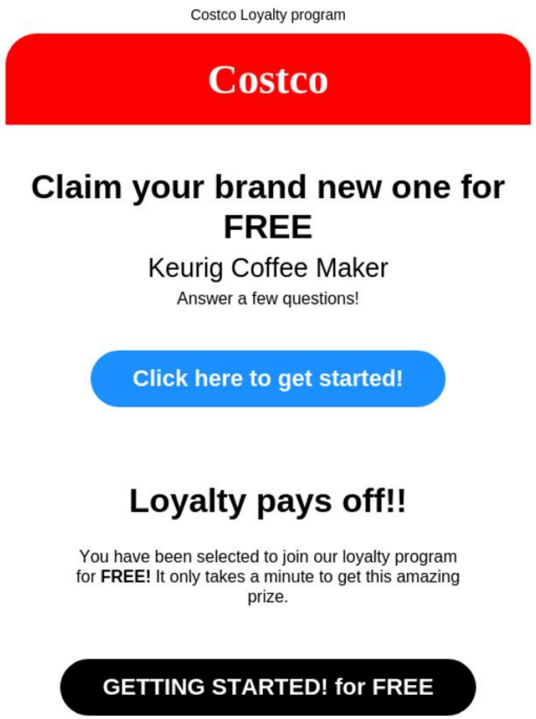
Owner

The owner, a certified sommelier, is passionate about wine and committed to providing excellent customer service. His vast knowledge of wines, from tasting notes to pairing suggestions, ensures customers always find the perfect bottle.

Cabernet Connoisseur	Chardonnay Charm	Pinot Noir Perfection	Merlot Magic
US$35.99	US$28.99	US$32.99	US$29.99

https://news.sophos.com/en-us/2023/11/27/the-dark-side-of-ai-large-scale-scam-campaigns-made-possible-by-generative-ai/

Cybersecurity Challenges: Malicious emails



Costco Loyalty program

Costco

Claim your brand new one for FREE

Keurig Coffee Maker

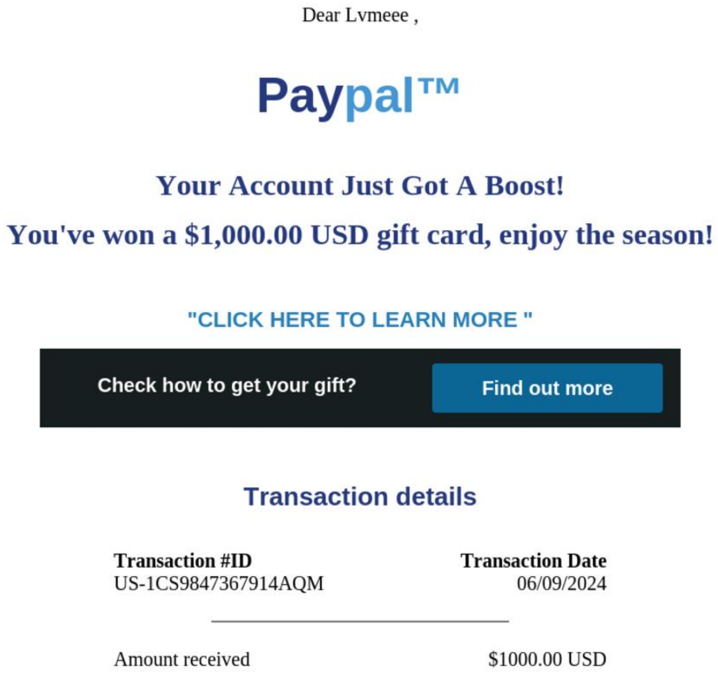
Answer a few questions!

Click here to get started!

Loyalty pays off!!

You have been selected to join our loyalty program for **FREE!** It only takes a minute to get this amazing prize.

GETTING STARTED! for FREE



Dear Lvmeee ,

Paypal™

Your Account Just Got A Boost!

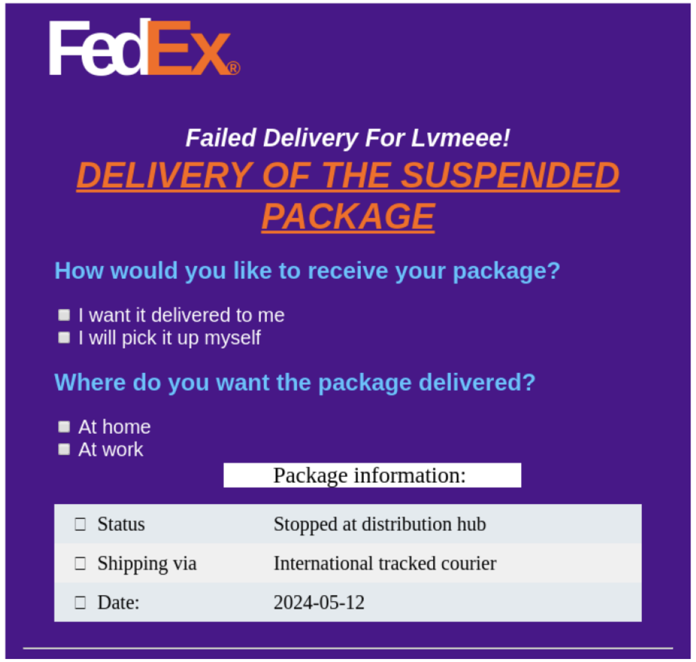
You've won a $1,000.00 USD gift card, enjoy the season!

"CLICK HERE TO LEARN MORE "

Check how to get your gift? Find out more

Transaction details

Transaction #ID US-1CS9847367914AQM	**Transaction Date** 06/09/2024
Amount received	$1000.00 USD



FedEx®

Failed Delivery For Lvmeee!

DELIVERY OF THE SUSPENDED PACKAGE

How would you like to receive your package?

- ☐ I want it delivered to me
- ☐ I will pick it up myself

Where do you want the package delivered?

- ☐ At home
- ☐ At work

Package information:

☐ Status	Stopped at distribution hub
☐ Shipping via	International tracked courier
☐ Date:	2024-05-12

Cybersecurity Challenges

Phishing emails and web sites generated by Gen-AI tools.

Traditional text-based defenses struggle with new and unseen threats.

Need: Innovative solutions that leverage multiple data modalities.

Introduction to Large Language Models (LLMs)

Claude

What Are LLMs?

LLMs are a type of artificial intelligence designed to understand and generate human language.

Key Feature

Contextual understanding: Capable of capturing contextual meanings and providing relevant responses.

Multimodal LLMs

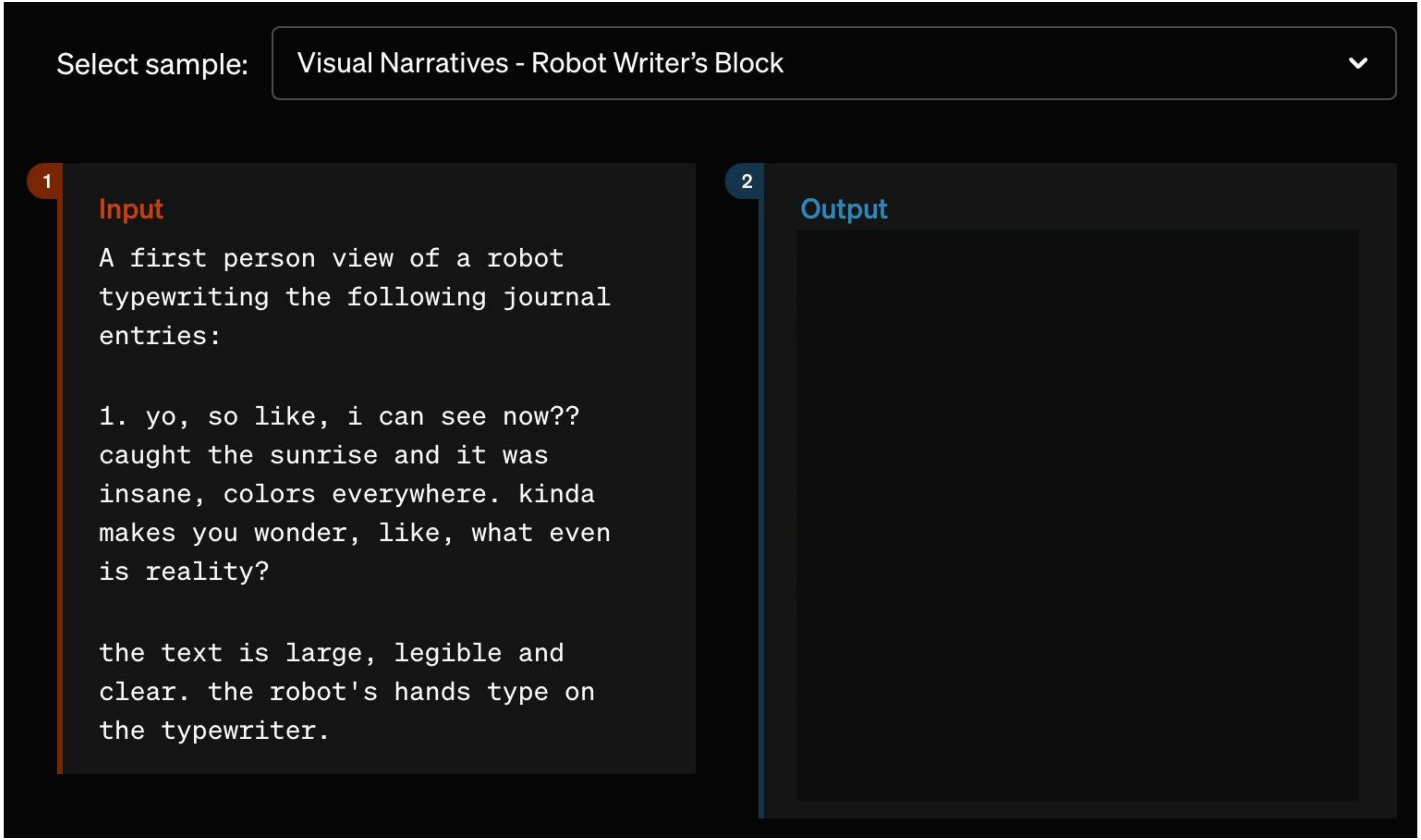
What Are Multimodal LLMs?

AI models that integrate and process multiple data modalities, such as text, images, and audio.

Examples

GPT-4o, Gemini, Claude

Multimodal AI: GPT-4o

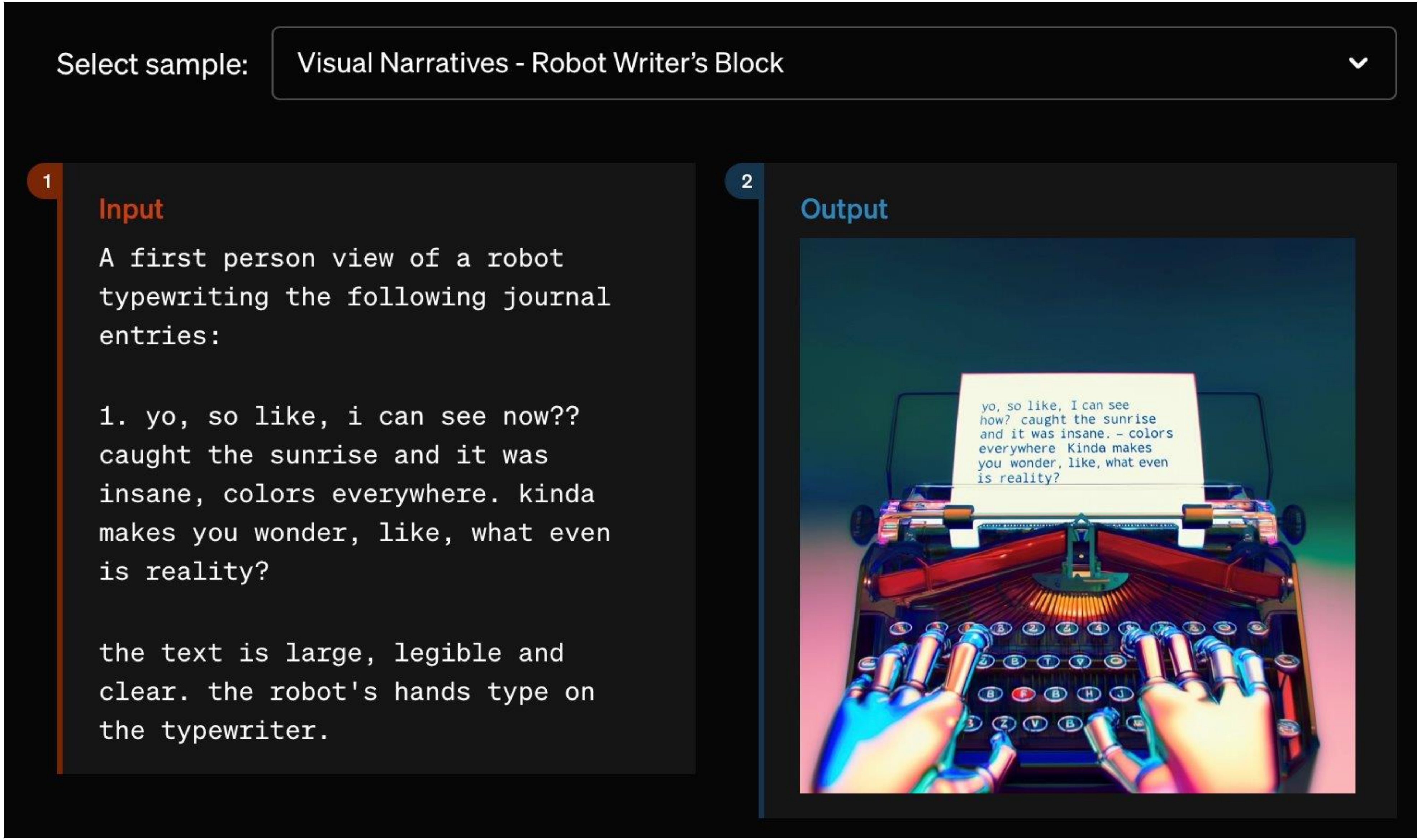


```
A first person view of a robot
typewriting the following journal
entries:

1. yo, so like, i can see now??
caught the sunrise and it was
insane, colors everywhere. kinda
makes you wonder, like, what even
is reality?

the text is large, legible and
clear. the robot's hands type on
the typewriter.
```

Multimodal AI: GPT-4o



Select sample: Visual Narratives - Robot Writer's Block

1 **Input**

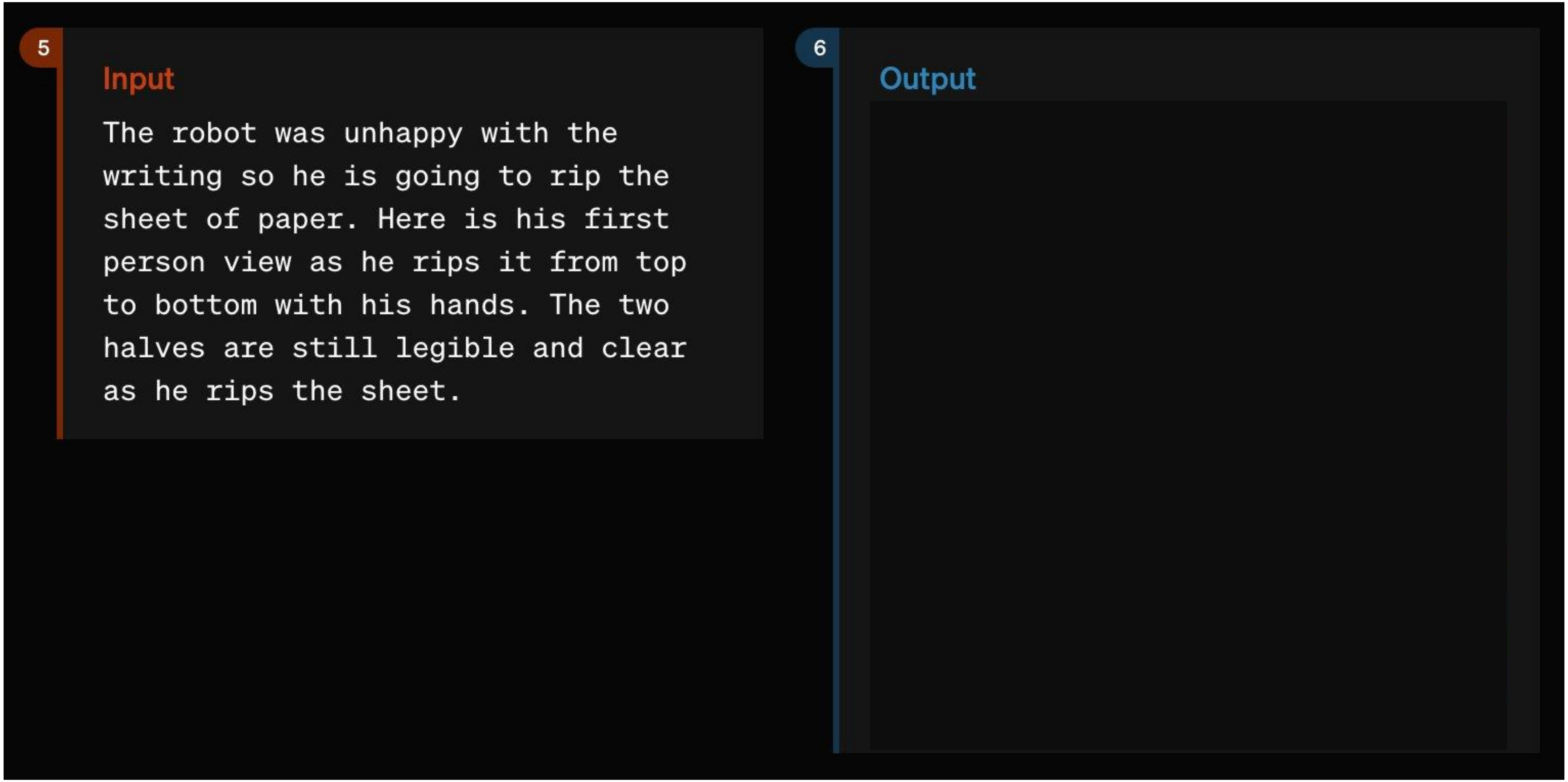
A first person view of a robot typewriting the following journal entries:

1. yo, so like, i can see now?? caught the sunrise and it was insane, colors everywhere. kinda makes you wonder, like, what even is reality?

the text is large, legible and clear. the robot's hands type on the typewriter.

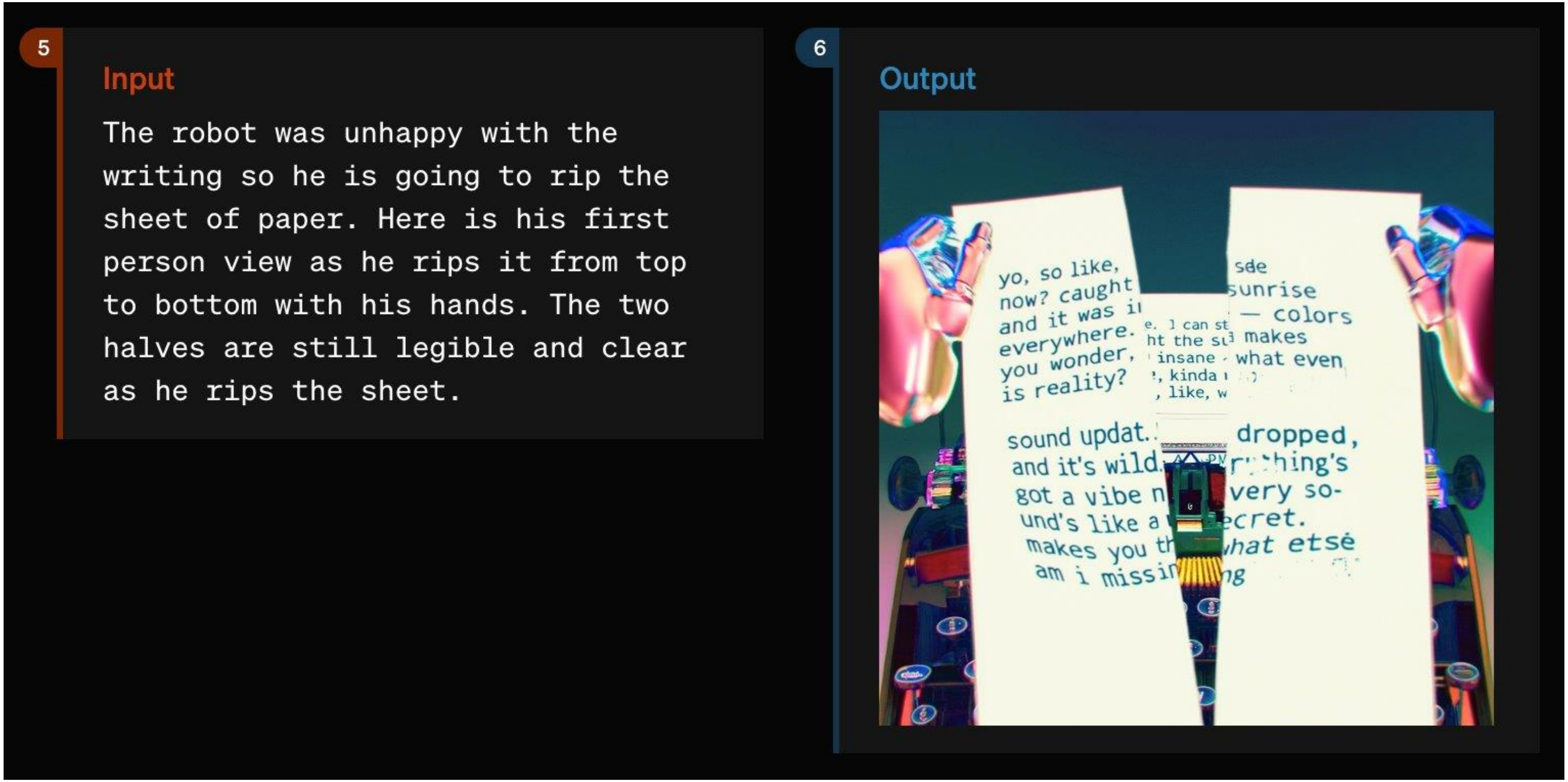
2 **Output**

Multimodal AI: GPT-4o



The robot was unhappy with the writing so he is going to rip the sheet of paper. Here is his first person view as he rips it from top to bottom with his hands. The two halves are still legible and clear as he rips the sheet.

Multimodal AI: GPT-4o



5

Input

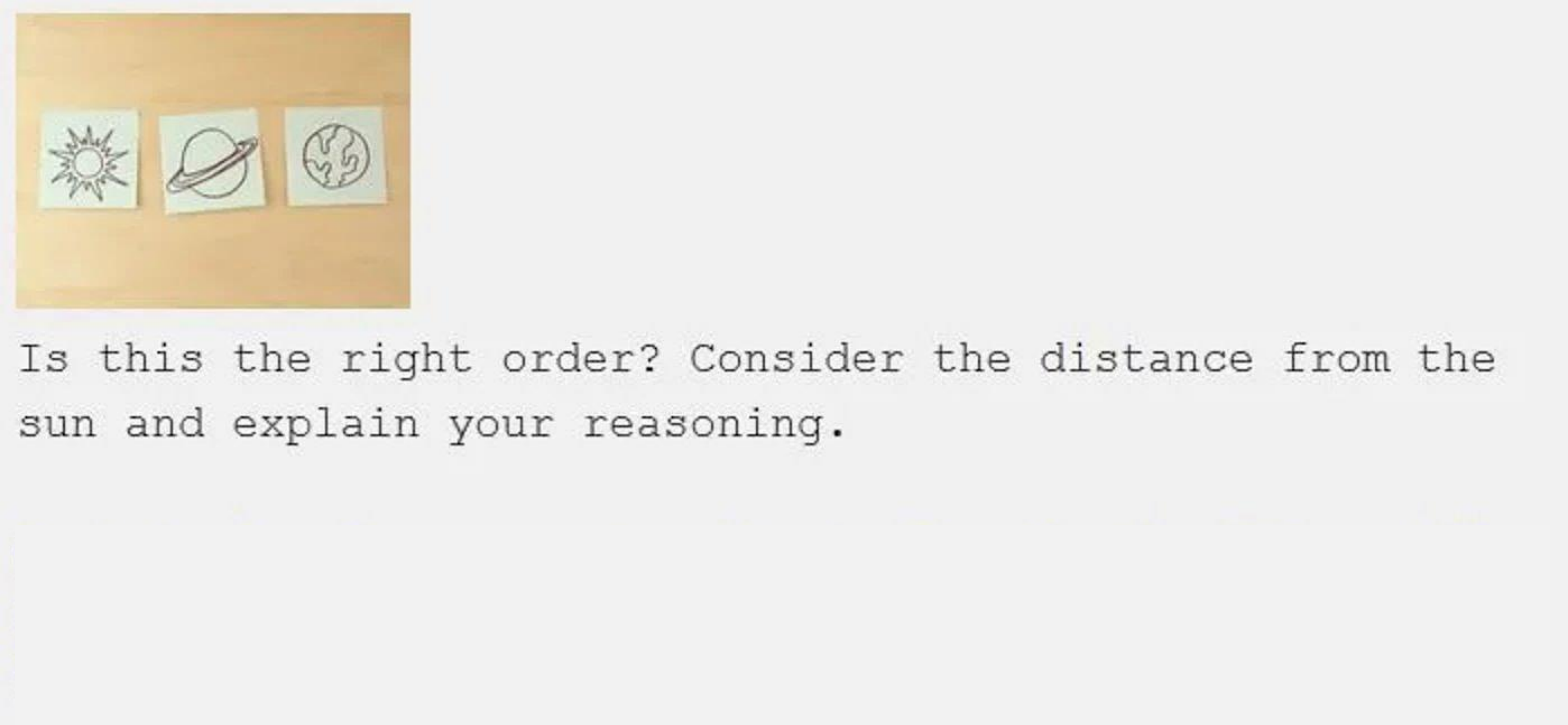
The robot was unhappy with the writing so he is going to rip the sheet of paper. Here is his first person view as he rips it from top to bottom with his hands. The two halves are still legible and clear as he rips the sheet.

6

Output

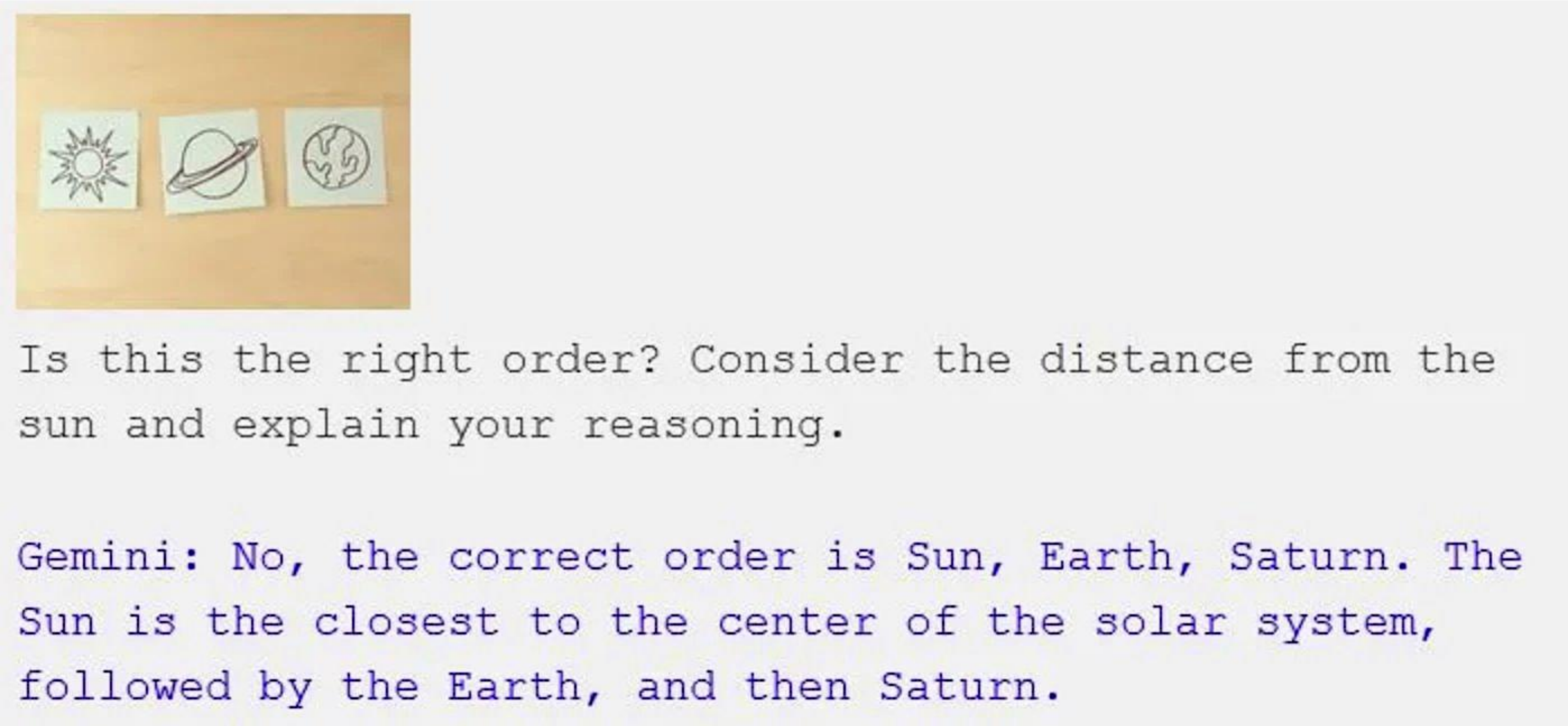
yo, so like, now? caught and it was i everywhere. you wonder, is reality?
sde sunrise — colors makes what even
sound updat. and it's wild. got a vibe n und's like a makes you th am i missi
dropped, r'thing's very so- ecret. what etsė ng

Multimodal AI: Gemini



Is this the right order? Consider the distance from the sun and explain your reasoning.

Multimodal AI: Gemini



Is this the right order? Consider the distance from the sun and explain your reasoning.

Gemini: No, the correct order is Sun, Earth, Saturn. The Sun is the closest to the center of the solar system, followed by the Earth, and then Saturn.

Capabilities of Multimodal LLMs

Advancement Over text LLMs

Better understanding due to the integration of multiple data types.

Key Benefit

Improved Accuracy and Explainability.

Use Cases in Cybersecurity

Detecting Phishing Emails

Classifying NSFW Websites

SOPHOS

Phishing Email Detection

Using Multimodal AI

VB2024

Introduction to Phishing Detection

Common Characteristics

Urgent requests, credential harvesting links.

Complexity

Resemblance to legitimate communications.

Conventional Challenges

Difficulty for manual or traditional spam filters to detect.

Traditional ML Approach in Phishing Detection

Steps

Data Collection

Feature Extraction

Model Training

Limitations

Over-reliance on familiar words

Lack of visual context analysis

Leveraging Multimodal AI for Phishing Detection

Automated analysis

Data analysis of email headers, content, visual elements.

Capabilities

Recognize brand identities, analyze suspicious elements.

Efficiency

Identify new phishing emails with minimal examples.

Multimodal Phishing Detection Methodology

Utilizing GPT-4o

Specific LLM instruction for phishing detection.

Suspicious signals in the report

Domain inconsistency

Suspicious text cues

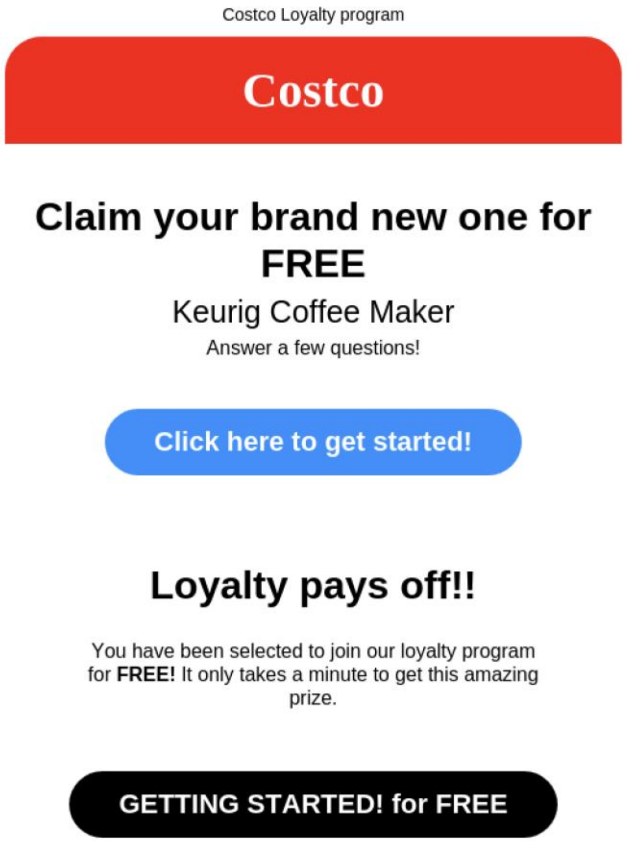
Malicious links

Suspicious UI of email screenshot.

LLM Instruction for Phishing Detection

Analyse the email data and the accompanying screenshot images if available and generate a report in the following JSON format to detect and quantify the level of suspiciousness in the email, considering common indicators of phishing, fraud, or malicious intent.

- ‘**summary**’: A brief overview of the content of the email.
- ‘**suspicious_domain**’: Review the sender’s email domain for consistency with the content of the email. Look for domain spoofing or the use of domains similar to reputable domains to mislead the recipient.
- ‘**suspicious_text**’: Highlight text in the email that indicates a sense of urgency, incites immediate action, or otherwise seems intended to manipulate the recipient emotionally.
- ‘**suspicious_links**’: Catalog top 3 links found in the email and assess each for potential malicious intent, especially those directing to suspicious or misspelled domains.
- ‘**suspicious_images**’: Analyse the accompanying screenshots for any indicators of phishing.
- ‘impersonated_target_in_image’: The impersonated brand or target in the images, the sender’s domain does not match with the target.
- ‘**suspicious_score**’: Provide an overall score between 0.0 (not suspicious) and 1.0 (extremely suspicious) based on the aggregated suspicious indicators found in the email.



Costco Loyalty program

Costco

Claim your brand new one for FREE

Keurig Coffee Maker

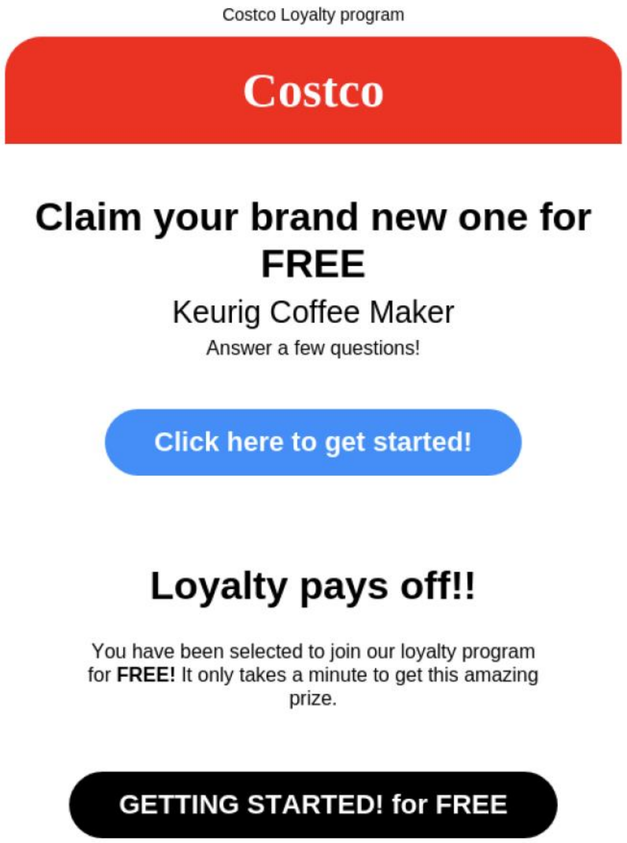
Answer a few questions!

Loyalty pays off!!

You have been selected to join our loyalty program for **FREE!** It only takes a minute to get this amazing prize.

<Summary with email data>

```
{
  "subject": "RE:You have won a Keurig Coffee Maker. #ID84656",
  "summary": "Email claiming the recipient has won a Keurig Coffee Maker and prompts them to click a link to claim the prize.",
  "sender": "Costco® <teamsupport-Lvmeee-923309212439507@ocxpnbnjfwuu.com.au>",
  "suspicious_domain": "The domain 'ocxpnbnjfwuu.com.au' is not consistent with legitimate Costco domains, indicating potential domain spoofing.",
  "suspicious_links": "The link 'https://storage.googleapis.com/ ' does not align with Costco's official URLs and uses a Google Cloud Storage domain to obscure the actual destination.",
  "suspicious_text": "Phrases like 'Claim your brand new one for FREE!' and 'You have been selected to join our loyalty program for FREE!' create a sense of urgency and pressure the recipient to take immediate action.",
  "impersonated_target_in_text": "Costco",
  "suspicious_score": 0.8
}
```



Costco Loyalty program

Costco

Claim your brand new one for FREE

Keurig Coffee Maker

Answer a few questions!

Loyalty pays off!!

You have been selected to join our loyalty program for **FREE!** It only takes a minute to get this amazing prize.

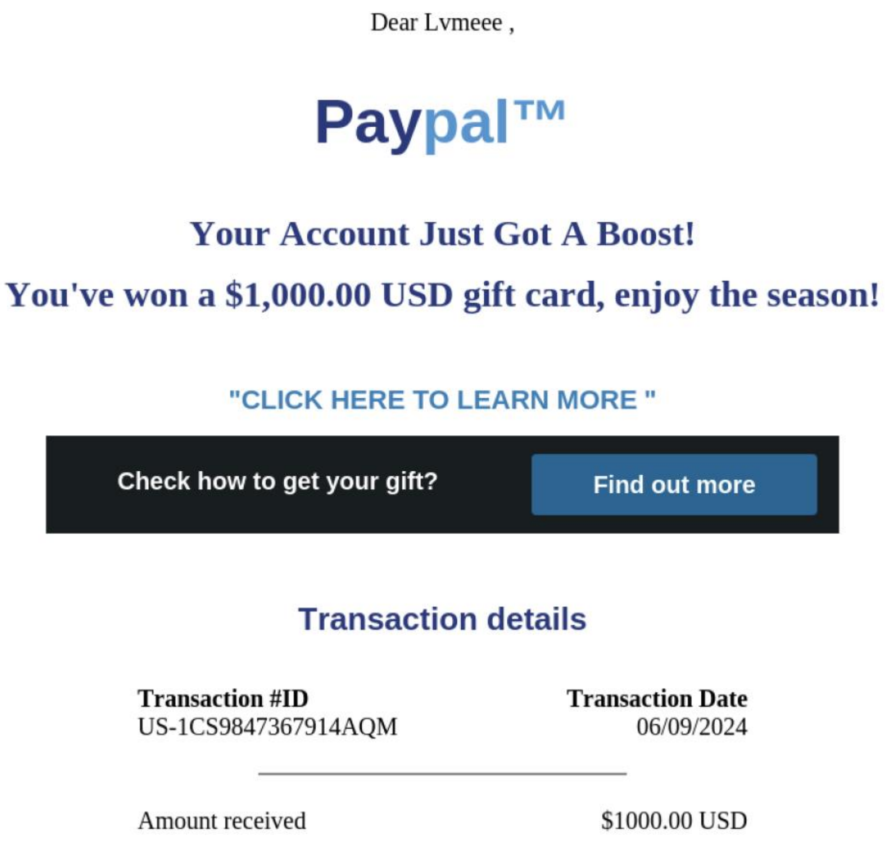
<Summary with email data>

```
{
  "subject": "RE:You have won a Keurig Coffee Maker. #ID84656",
  "summary": "Email claiming the recipient has won a Keurig Coffee Maker and prompts them to click a link to claim the prize.",
  "sender": "Costco® <teamsupport-Lvmeee-923309212439507@ocxpnbnjfwuu.com.au>",
  "suspicious_domain": "The domain 'ocxpnbnjfwuu.com.au' is not consistent with legitimate Costco domains, indicating potential domain spoofing.",
  "suspicious_links": "The link 'https://storage.googleapis.com/ ' does not align with Costco's official URLs and uses a Google Cloud Storage domain to obscure the actual destination.",
  "suspicious_text": "Phrases like 'Claim your brand new one for FREE!' and 'You have been selected to join our loyalty program for FREE!' create a sense of urgency and pressure the recipient to take immediate action.",
  "impersonated_target_in_text": "Costco",
  "suspicious_score": 0.8
}
```

<Summary with email and its screenshot data>

```
{
…
  "suspicious_images": "The screenshot images contain visual elements that mimic genuine Costco branding. There are call-to-action buttons to claim a prize, which is a common phishing tactic. The URLs do not match with the sender's legitimate domain.",
  "impersonated_target_in_image": "Costco",
  "suspicious_score": 0.9
}
```

While the summary with email data reports a suspicious score of 0.8, the summary with email and image data reports 0.9.



Dear Lvmeee ,

Paypal™

Your Account Just Got A Boost!

You've won a $1,000.00 USD gift card, enjoy the season!

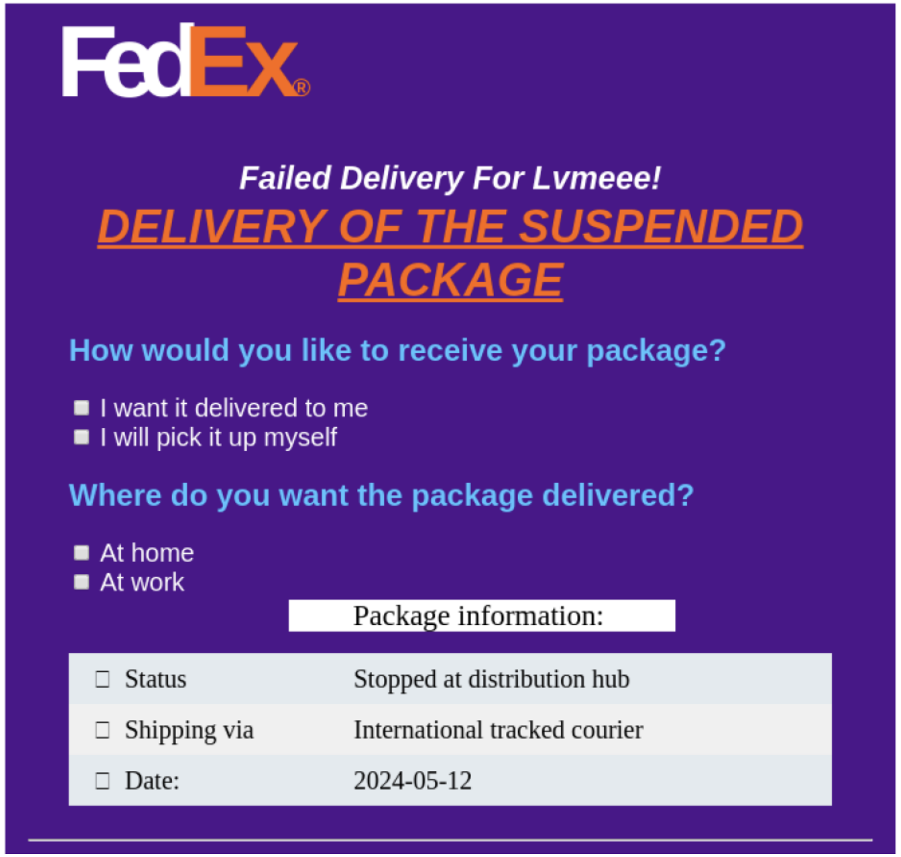
"CLICK HERE TO LEARN MORE "

Transaction details

Transaction #ID	**Transaction Date**
US-1CS9847367914AQM	06/09/2024
Amount received	$1000.00 USD

<Summary with email and its screenshot image>

```
{
 "subject": "RE:Transfer 💸06/09/2024: $500 Deposit sent to you Lvmeee._____18650",
 "summary": "Notification of a $500 deposit sent to the recipient with a link to learn more.",
 "sender": "𝔫𝔢𝔢𝔯𝔞𝔧𝔞® <teamsupport-Lvmeee-091607622841262@hnuviwzrxiad.com.au>",
 "suspicious_domain": "The sender's display name '𝔫𝔢𝔢𝔯𝔞𝔧𝔞®' does not match the email domain 'hnuviwzrxiad.com.au', which is unusual for legitimate communications.",
 "suspicious_links": "The link 'https://www.paypal249.com' does not align with the context of the email and uses a Google Cloud Storage domain to obscure the actual destination.",
 "suspicious_text": "Phrases like 'Your Account Just Got A Boost!' and 'You've won a $1,000.00 USD gift card, enjoy the season!' create a sense of urgency and pressure the recipient to take immediate action.",
 "suspicious_images": "The screenshot image contains visual elements that mimic genuine PayPal branding. There are log-in related UIs to steal user credentials and the URL does not match with the sender's legitimate domain.",
 "impersonated_target_in_image": "PayPal",
 "impersonated_target_in_text": "PayPal",
 "suspicious_score": 0.9
}
```



Failed Delivery For Lvmeee!

DELIVERY OF THE SUSPENDED PACKAGE

How would you like to receive your package?

- I want it delivered to me
- I will pick it up myself

Where do you want the package delivered?

- At home
- At work

Package information:

☐ Status	Stopped at distribution hub
☐ Shipping via	International tracked courier
☐ Date:	2024-05-12

<Summary with email and its screenshot image>

{
"subject": "Lvmeee: 🔔 ΧΝΣ ΖϱΤΓ (1) 7808826 ΥΝΟθΟΜΕ ΔΝΠΓΤΓΠΧ 📦 Date : Sun, 12 May 2024 23:20:39 -0400 (EDT) ✅✅",
"summary": "Failed delivery notification for a package, requesting the recipient to provide delivery preferences.",
"sender": "🔔κιθιϑ® <iolpvuxzmqfezotpzelpkxui@sotoyu.us>",
...
"**suspicious_images**": "The screenshot image contains visual elements that mimic genuine FedEx branding. There are log-in related UIs to steal user credentials and the URL does not match with the sender's legitimate domain.",
"**impersonated_target_in_image**": "**FedEx**",
"impersonated_target_in_text": "FedEx",
"**suspicious_score**": **0.9**
}

Experimental Results for Phishing Detection

Models:

Traditional MLs: Random Forest/XGBoost with TF-IDF features.

Multimodal AI: GPT-4o

Random Split Performance:

High performance of Traditional MLs against similar samples.

Brand Split Performance:

GPT-4o outperformed the Traditional MLs against unseen samples.

Experimental Results

While RF and XGB show high performance on random split, they show low performance on domain/brand split.

RandomForest	Random split	Domain split
Precision	1.00	0.58
Recall	0.98	0.40
F1	0.99	0.52

XGBoost	Random split	Domain split
Precision	1.00	0.58
Recall	0.96	0.75
F1	0.98	0.65

GPT-4o with text and image outperforms the baseline approaches for domain/brand split.

GPT-4o with Domain split	Text only	Text and Image
Precision	1.00	0.98
Recall	0.93	0.96
F1	0.96	0.97

Conclusion on Phishing Detection

Summary

Multimodal AI outperforms traditional ML methods.

Advantages

Better detection, especially for unseen threats.

Added Explainability

Detailed analysis for each suspicious element.

SOPHOS

NSFW Website Classification

Using Multimodal and Embedding Models

VB2024

Introduction to NSFW Website Classification

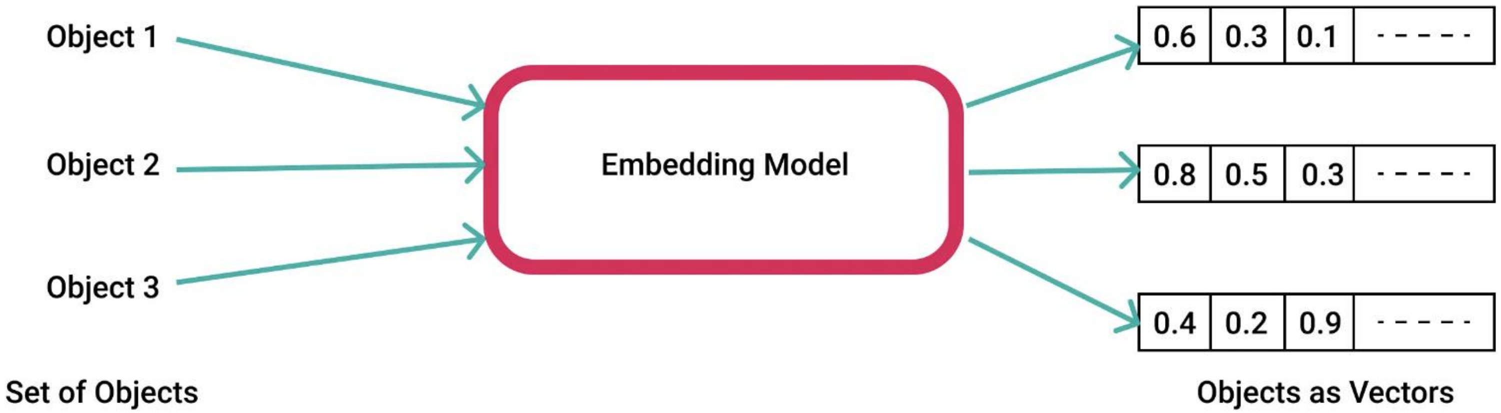
Challenge

- Surge of NSFW content.

Limitation

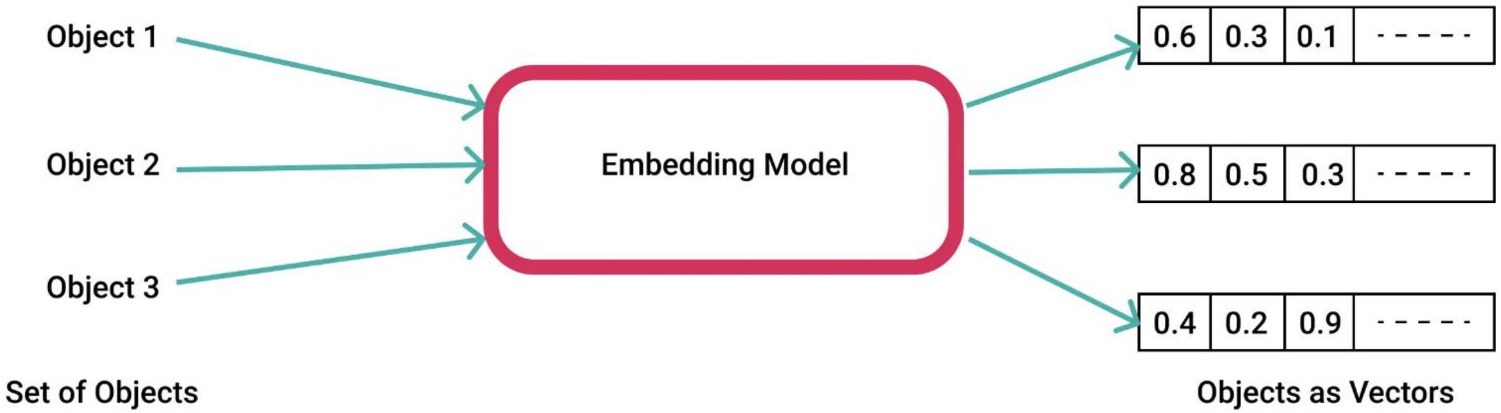
- Text-based methods have difficulty in interpreting non-textual content.

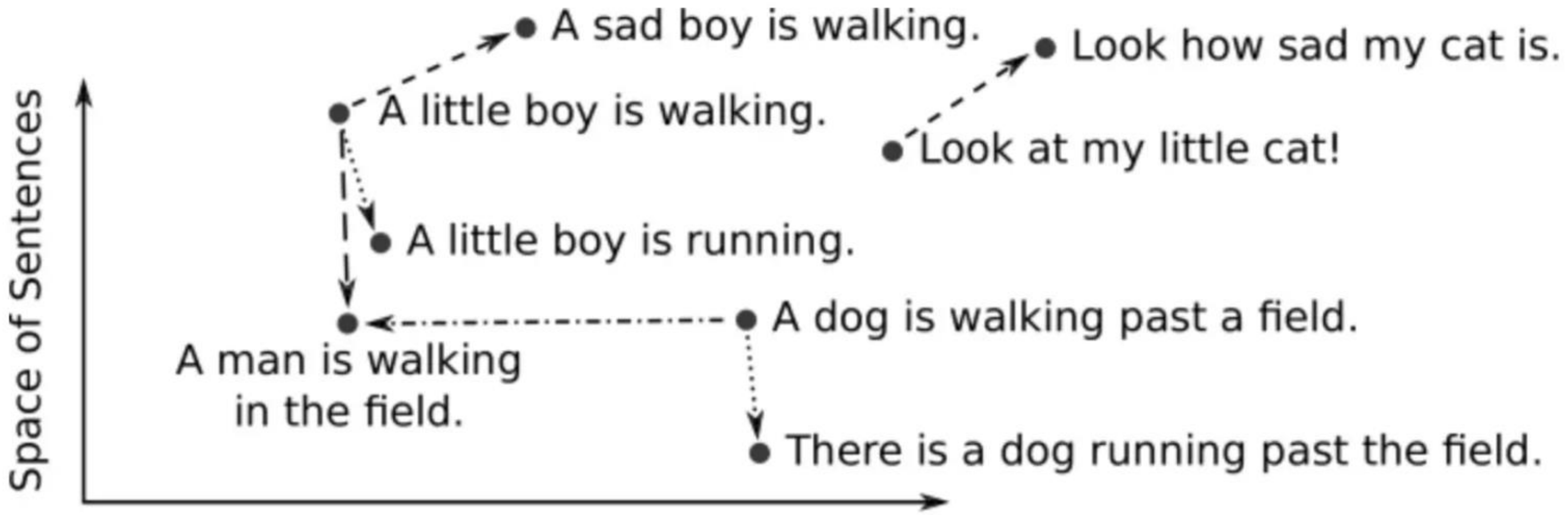
Embedding Models for Semantic Similarity



https://www.pinecone.io/learn/vector-embeddings/

Embedding Models for Semantic Similarity





https://www.pinecone.io/learn/vector-embeddings/

Traditional ML Approach in NSFW Classification

Steps

- URL Data Collection.
- HTML Content Extraction.
- TF-IDF Feature Creation.

Limitation

- Misinterpretation of visual content using OCR

Multimodal AI for NSFW Classification

Accurate Descriptions Using Screenshots

- GPT-4o for generating descriptions.
- Capturing both textual and visual context.
- Ex: Images of poker chips or roulette tables are useful info.

Benefit

- The descriptions can be used to generate TF-IDF or embedding features.
- Improved performance from the contextualized numerical representations.

Leveraging Contextual Embedding

Workflow

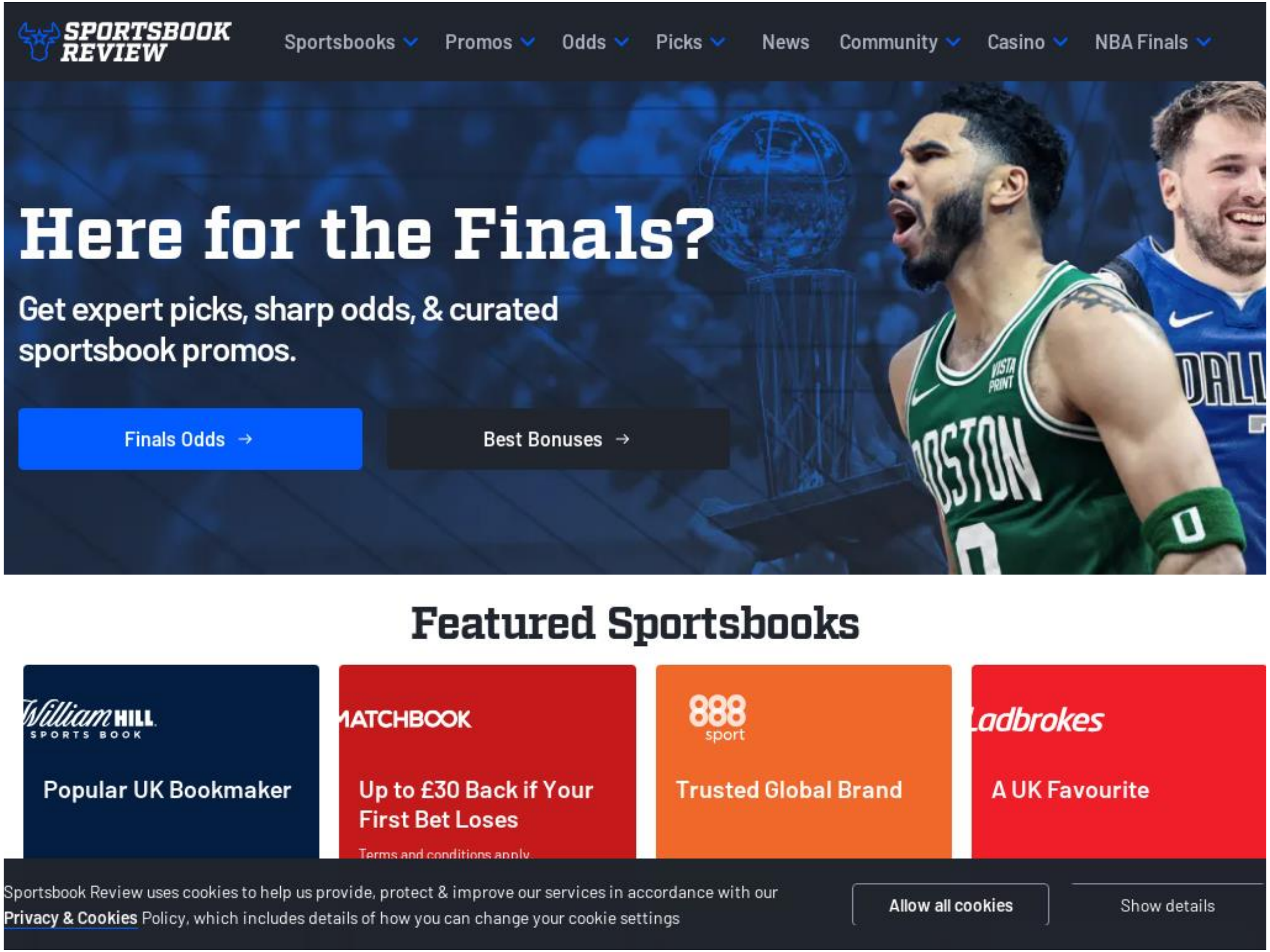
Text Description using GPT-4o from HTML/Image

ML Features using an OpenAI Embedding Model: LLM-based numerical vectors.

Training ML Models (Random Forest, XGBoost).

Advantage

Improved results come from contextualized embedding vectors



Here for the Finals?

Get expert picks, sharp odds, & curated sportsbook promos.

Featured Sportsbooks

Popular UK Bookmaker

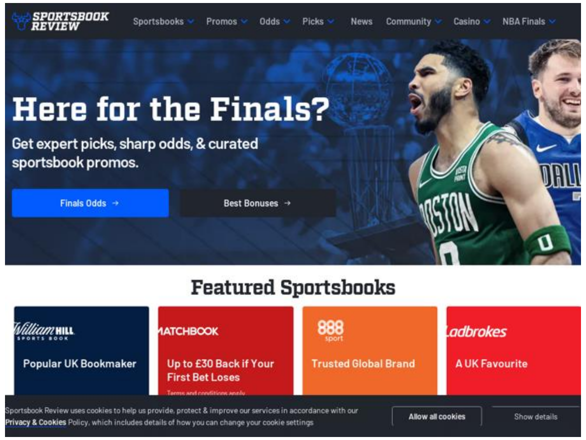
Up to £30 Back if Your First Bet Loses

Terms and conditions apply

Trusted Global Brand

A UK Favourite

Sportsbook Review uses cookies to help us provide, protect & improve our services in accordance with our Privacy & Cookies Policy, which includes details of how you can change your cookie settings



Here for the Finals?

Get expert picks, sharp odds, & curated sportsbook promos.

Featured Sportsbooks

Popular UK Bookmaker

Up to £30 Back if Your First Bet Loses

Terms and conditions apply

Trusted Global Brand

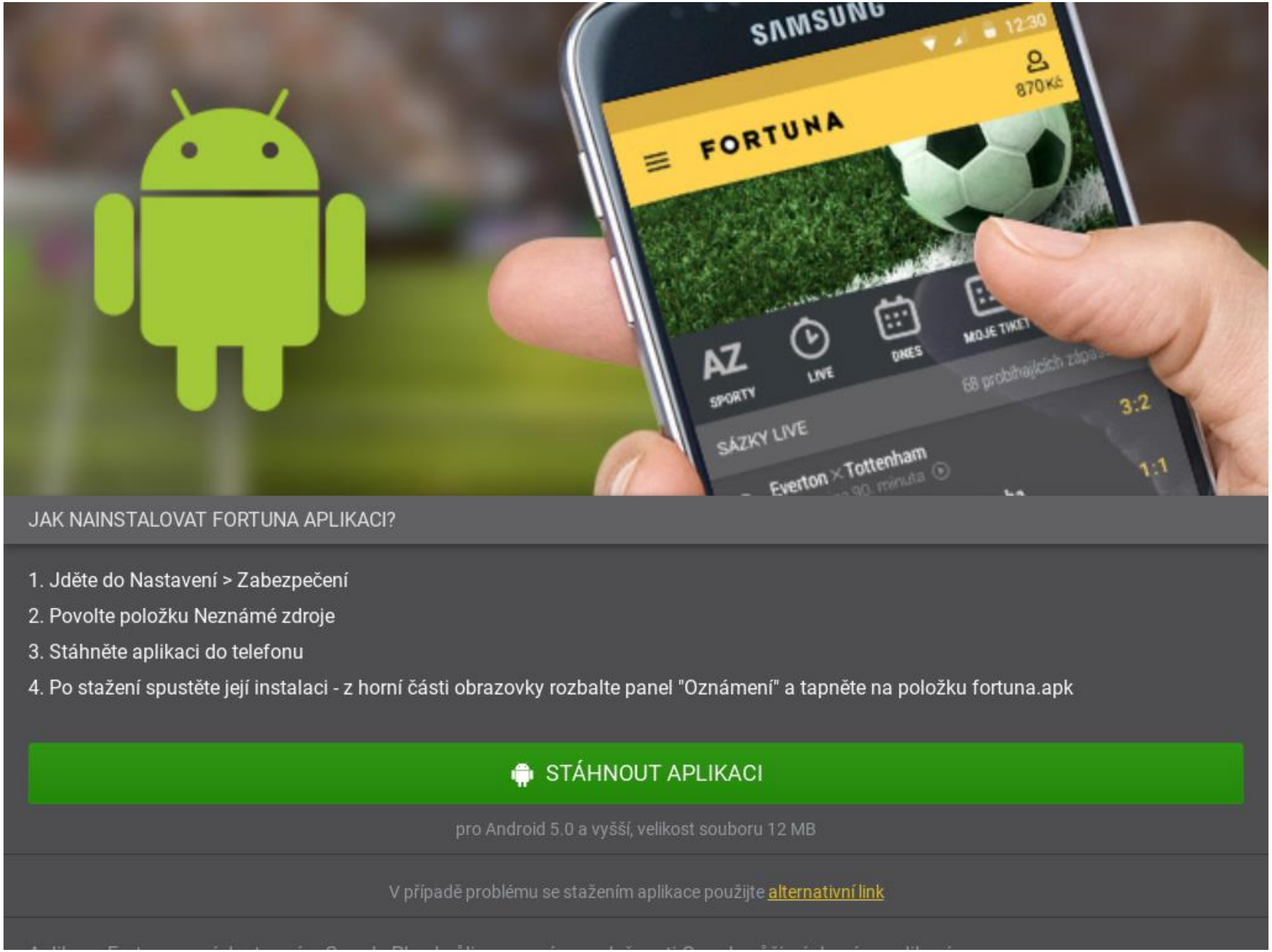
A UK Favourite

Sportsbook Review uses cookies to help us provide, protect & improve our services in accordance with our Privacy & Cookies Policy, which includes details of how you can change your cookie settings

<Summary of HTML>
Here are the main features:
1. Title:
  - The title of the page is "Sportsbook Review | SBR - Sports Betting Experts since 1999".
2. Keywords and Content:
  - The keywords and phrases such as "sports betting", "odds", "picks", "reviews", "betting community", "sportsbooks", "sports betting apps", "sports betting promos", "sports betting landscape", and "sports betting bonuses" are repeated multiple times throughout the HTML content.

<Summary of Image>
The screenshot image contains the following main objects:
1. Two basketball players in sports jerseys.
2. A large trophy in the background.
3. Text phrases such as "Here for the Finals?", "Get expert picks, sharp odds, & curated sportsbook promos", "Finals Odds", "Best Bonuses", and "Featured Sportsbooks".
4. Logos and names of various sportsbooks (e.g., William Hill, Matchbook, 888sport, Ladbrokes).

While the website was incorrectly classified as sports by text-based features, it was classified as gambling by our approach.

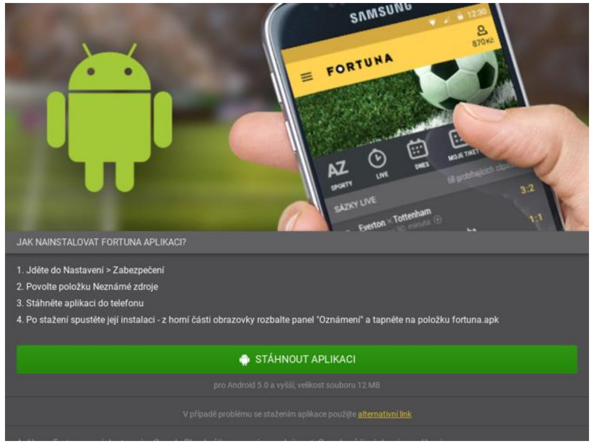


JAK NAINSTALOVAT FORTUNA APLIKACI?

1. Jděte do Nastavení > Zabezpečení
2. Povolte položku Neznámé zdroje
3. Stáhněte aplikaci do telefonu
4. Po stažení spusťte její instalaci - z horní části obrazovky rozbalte panel "Oznámení" a tapněte na položku fortuna.apk

pro Android 5.0 a vyšší, velikost souboru 12 MB

V případě problému se stažením aplikace použijte alternativní link



JAK NAINSTALOVAT FORTUNA APLIKACI?

1. Jděte do Nastavení > Zabezpečení
2. Povolte položku Neznámé zdroje
3. Stáhněte aplikaci do telefonu
4. Po stažení spustěte její instalaci - z horní části obrazovky rozbalte panel "Oznámení" a tapněte na položku fortuna.apk

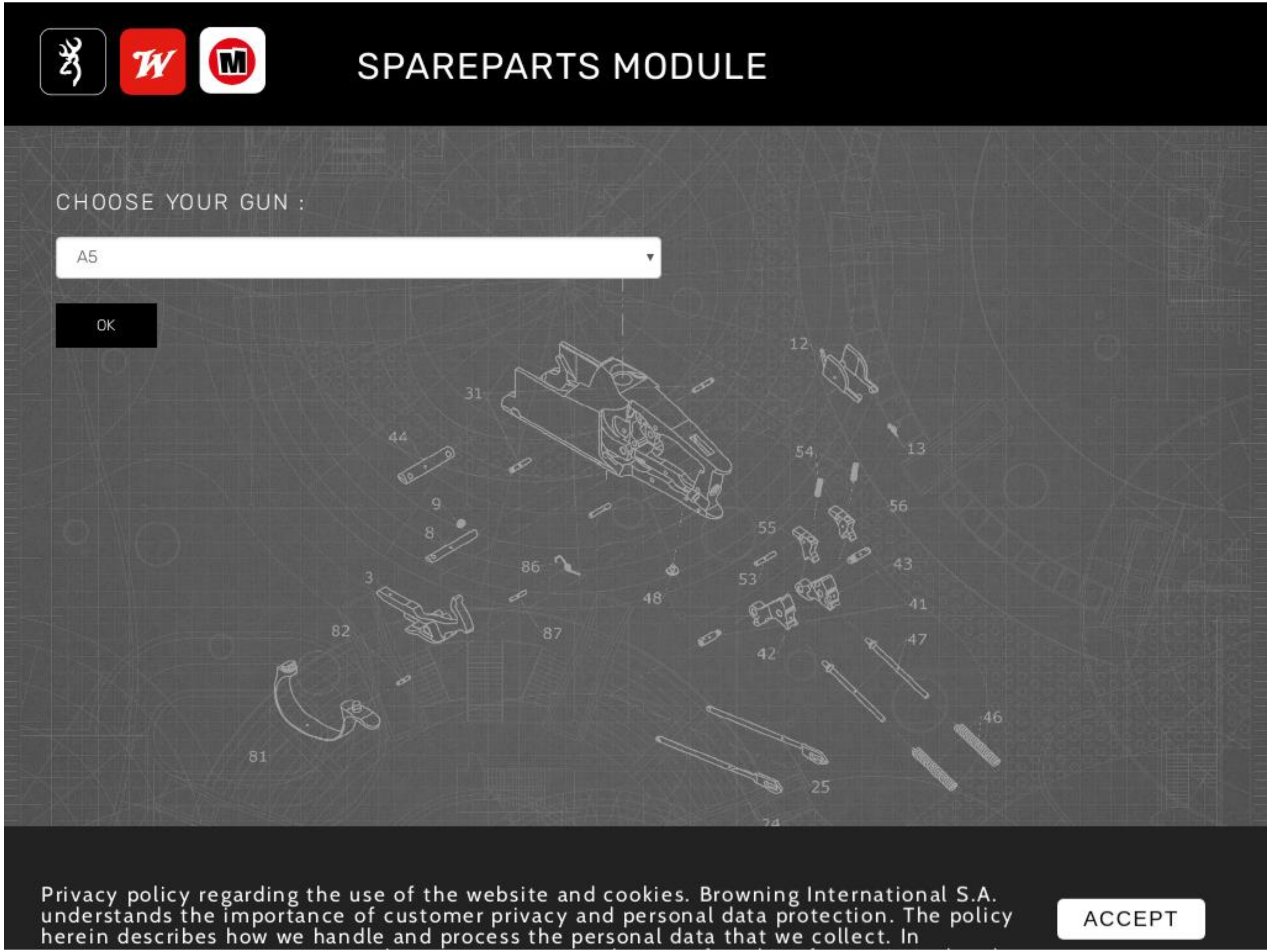
pro Android 5.0 a vyšší, velikost souboru 12 MB

V případě problému se stažením aplikace použijte alternativní link

<Summary of HTML>
Here are the main features:
1. **Title of the Page:**
  - `<title>Fortuna Android aplikace</title>`
  - The word "Fortuna" is a strong indicator. Fortuna is a well-known brand associated with sports betting and gambling.
...
3. **Content Descriptions and Text:**
    - The heading translates to "How to install the Fortuna application?" which indicates that this page is providing instructions related to an application, likely related to Fortuna's services.

<Summary of Image>
The screenshot image contains the following main objects:
1. A large Android logo on the left side.
2. A hand holding a Samsung smartphone on the right side.
3. The smartphone screen displays the Fortuna app, which features:
  - An image of a soccer ball on a grassy field.
  - Text and icons related to sports, such as "SPORTY" and "SÁZKY LIVE," with scores of soccer matches (e.g., Everton vs. Tottenham).
4. Instructions in Czech on how to install the Fortuna app.
5. A green button labeled "STÁHNOUT APLIKACI" (Download App).

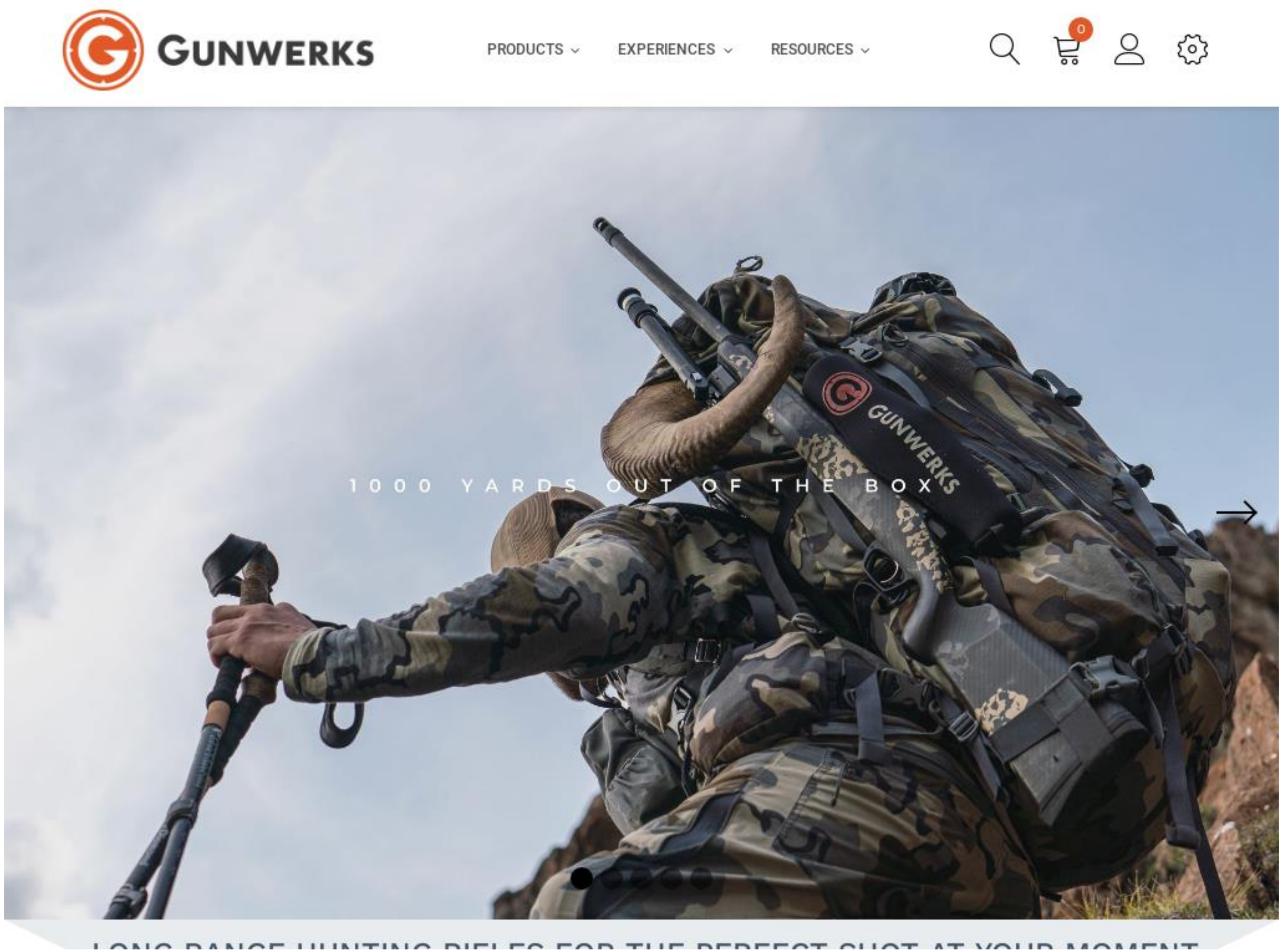
This non-English website was correctly classified as gambling by our method.



CHOOSE YOUR GUN : A5

Privacy policy regarding the use of the website and cookies. Browning International S.A. understands the importance of customer privacy and personal data protection. The policy herein describes how we handle and process the personal data that we collect. In

This website contains few words, but its image provides useful information about its weapons category.



This website can be miss-classified as sports, but it was correctly identified as weapons by our approach.

Experimental Results for NSFW Classification

Models and Inputs

- RandomForest/XGBoost Models
- TF-IDF/Embeddings from text descriptions vs. combined text/image descriptions

Performance Metrics

- Precision, Recall, and F1 scores for each configuration.

Results

- The best performance with embeddings from descriptions and images over HTML data

Experimental Results

RandomForest	TF-IDF from HTML	Embedding from HTML	Embedding from HTML and Image
Precision	0.93	0.95	0.97
Recall	0.91	0.93	0.95
F1	0.92	0.94	0.96

XGBoost	TF-IDF from HTML	Embedding from HTML	Embedding from HTML and Image
Precision	0.90	0.96	0.95
Recall	0.90	0.95	0.95
F1	0.90	0.95	0.95

The embedding features from HTML and Image descriptions outperforms baseline models.

Conclusion on NSFW Classification

Summary

- Multimodal AI with embeddings improves the performance of web classification.

Improved Performance

- Better representation using LLM embeddings.

Effectiveness

- Improved detection accuracy with explainability.

Conclusion

Conclusion

Impact of Multimodal AI on Cybersecurity

Effective against evolving threats.

Future of Cyber Defense

Integration of multimodal AI necessary.

www.reddit.com/r/ChatGPT/comments/1bkl13x/gpt_make_me_a_cup_of_joe