Ben-Gurion University of the Negev

Faculty of Engineering Science

School of Electrical and Computer Engineering

Dept. of Electrical and Computer Engineering

Fourth Year Engineering Project

Final Report

MemeMend: Elevating Political Discourse

Project number:	P2025-067
Students	
(name & ID):	Bar Binyamin Varsulker Roy Ayalon
Supervisors:	Dan Vilenchik
Sponsors:	
Submitting date:	20/07/2025

MemeMend משפרים את השיח הפוליטי

וורסולקר בר בנימין, אילון רועי
varsulba@post.bgu.ac.il
וילנצ'יק דן

בעידן הדיגיטלי, ממים הם כלי תקשורת פופולרי להעברת מסרים, אך חלקם מכילים תכנים פוגעניים העלולים לגרום לפגיעה רגשית בקרב משתמשים שונים .מטרת הפרויקט היא לפתח מערכת לזיהוי ממים בעלי תוכן פוגעני ברמת דיוק גבוהה ולהציע תחליפים שאינם פוגעניים תוך שמירה על כוונת המסר המקורית. לשם סיווג ממים, חילצנו את הטקסט באמצעות GOT_OCR2 , והעברנו כל רכיב (טקסט, תמונה, מם) למודלי סיווג ייעודיים. תוצאת OR קבעה את סיווג המם. ממים פוגעניים נותחו ב Gamma-3 שיצר הסבר ומם חלופי (תיאור תמונה וטקסט). תיאור התמונה הוזן ל- Stable Diffusion שיצר את התמונה שהטקסט הודבק עליה ליצירת מם חדש. לשם ניתוח המערכת יצרנו דאטהסט הכולל 1,000 ממים, שכחצי מהם פוגעניים. סיווג הטקסט והתמונה השיגו כל אחד דיוק של 69%, סיווג המם 72%, ושילוב שלושתם הניב דיוק של 75%. לסיכום, המערכת מצליחה לזהות ממים פוגעניים בדיוק גבוה ולהציע אלטרנטיבה מתאימה, תוך שמירה על מסר המם המקורי. מילות מפתח: ממים, סינון תוכן, העברת מסרים, רשתות נוירונים, בינה מלאכותית, רגישות תרבותית, ניתוח טקסט ותמונה.

MemeMend: Elevating Political Discourse

Varsulker Bar Binyamin, Ayalon Roy
varsulba@post.bgu.ac.il
Vilenchik Dan

In the digital age, memes have become a popular communication tool for conveying messages, but some contain offensive content that may cause emotional harm to various users. The goal of this project is to develop a system that accurately detects offensive memes and offers non-offensive alternatives while preserving the original intent of the message. To classify memes, we extracted the text using GOT_OCR2 and passed each component (text, image, meme) through dedicated classification models. An OR operation on the results determined the meme's final classification. Offensive memes were analyzed by Gamma-3, which generated both an explanation and an alternative meme (image description and text). The image description was fed into Stable Diffusion to generate a new image, onto which the alternative text was added to create a new meme. For system evaluation, we built a dataset of 1,000

manually labeled memes, approximately half of which were offensive. Text and image classification each achieved 69% accuracy, meme classification achieved 72%, and the combined model reached 75%.
In conclusion, the system successfully detects offensive memes with high accuracy and suggests suitable alternatives that preserve the original message.
Keywords: memes, content filtering, message transmission, neural networks, artificial intelligence, cultural sensitivity, text and image analysis.

MemeMend: Elevating Political Discourse

In recent years, memes have become a dominant form of communication on the internet, often blending humor, culture, and commentary in a single image. However, the same expressive power that makes memes so impactful also enables them to be used as tools for spreading offensive, harmful, or discriminatory content. As the prevalence of such content grows, so does the need for intelligent systems that can automatically detect and mitigate the spread of offensive memes while preserving the freedom of expression that memes allow. The motivation for this project stems from the need to create a balanced solution that preserves free expression while combating the spread of harmful content. While previous work, such as Facebook's Hateful Memes Challenge [1], has explored meme classification, these efforts were limited to labeling memes as hateful or not. To the best of our knowledge, no existing system offers an end-to-end solution that not only classifies offensive memes but also:

1. Explains why a meme is considered offensive.
2. Generates a non-offensive alternative that preserves the original message and meme-like qualities.

The objective of this project is to develop a comprehensive system capable of:

1. Detecting whether a meme is offensive.
2. Explaining the reasoning behind this classification.
3. Generating a revised, non-offensive meme that retains the original intent and tone.
4. Exposing the entire pipeline through a user-friendly API, making the solution easily integrable into platforms, tools, or moderation systems.

   Due to hardware constraints and the absence of a large, labeled dataset of offensive and non-offensive memes, we adopted a modular approach that chains together existing pre-trained models. This allows us to leverage cutting-edge tools in OCR, vision-language understanding, text-to-image generation, and natural language processing, while optimizing for efficiency and scalability. The resulting pipeline is designed to be both practical and extensible, offering a novel contribution to the field of automated content moderation.

Figure 1: Examples Of Memes

Technical Specification

System Overview

The system is composed of modular components, each responsible for a specific task in the offensive meme detection and transformation pipeline. The full process is illustrated in the following block diagram

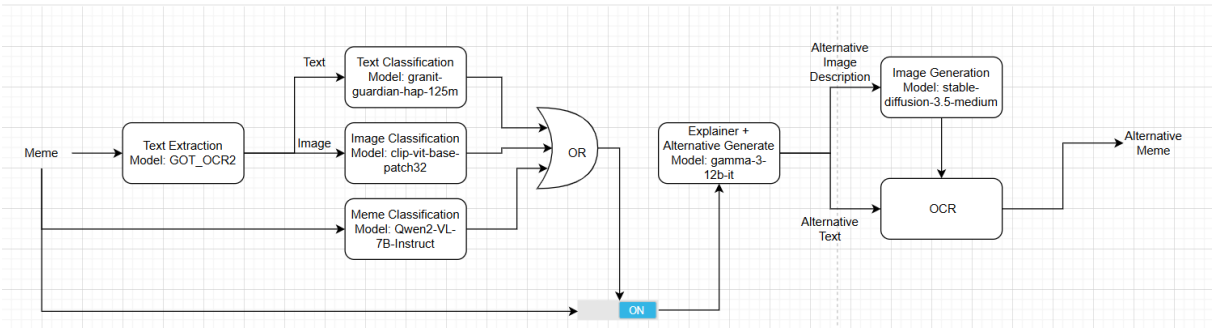


Figure 2: System Architecture Diagram

The pipeline starts with a meme image as input and proceeds through several stages, including text extraction, multimodal classification, explanation and regeneration, and final assembly of a non-offensive meme. The system is exposed through a user-friendly chatbot, enabling easy integration into third-party platforms or tools.

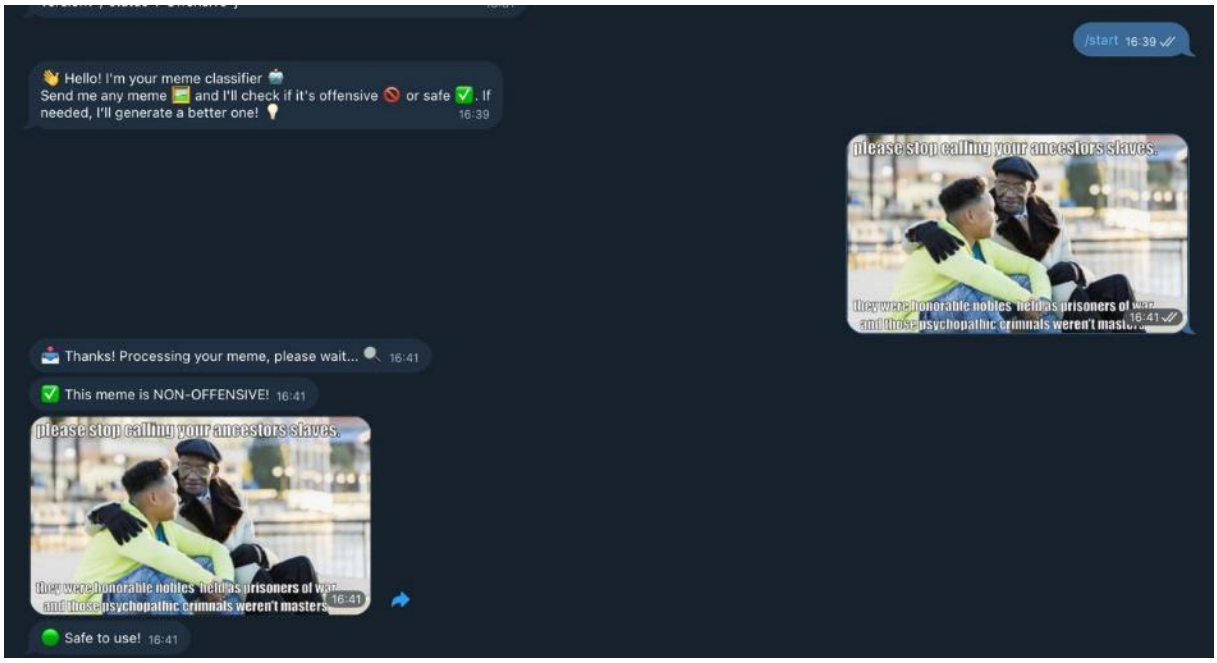


Figure 3: User-Friendly Chatbot

Hardware Requirements

CPU: Intel i7 or equivalent

GPU: 4xNVIDIA A5000 or higher with minimum 24GB VRAM

RAM: Minimum 16GB

Storage: ~100GB for model weights and logs.

OS: Ubuntu 20.05 / Windows 11.

Programming language: Python 3.10 or higher.

Module Description

1. Text Extraction:
   Model: GOT_OCR2.0 [2]
   Input: Meme image
   Output: Extracted overlay text
   Used to separate textual content from the meme image for analysts.
2. Offensiveness Classification
   Text Classification Model: granit-guardian-hap-125m [3]
   Image Classification Model: clip-vit-base-patch32 [4]
   Meme Classification Model: Qwen2-VL-7B-Instruct fine-tuned [5]
   Classification outputs feed into a conditional switch logic (OR gate) to decide offensiveness.
3. Explanation and Generation

Model: gemma-3-4b-it [6]

If the meme is classified as offensive, the model provides a natural-language explanation of the offensiveness, generates alternative meme content (text and description) that retains the original message.

4. Image Generation

   Model: stable-diffusion-3.5-medium [7]

   Takes the alternative image description and generates a new meme background image.

5. Final Composition

   The alternative text is overlayed on the generated image using an OCR-aligned placement tool, producing the alternative meme.

The modular structure and lightweight model selection enable efficient operation even under hardware constraints, making the system suitable for academic deployment and scalable real-world use.

Performance

To evaluate the accuracy of the classification system, we used a dataset from the Facebook Toxicity Challenge containing 10,000 memes classified as either toxic or non-toxic. After manually reviewing the dataset, we found that many memes were incorrectly labeled, which would affect the reliability of the results. Therefore, we created a smaller, manually curated, and more reliable dataset of 1,000 memes, approximately half of which were labeled as toxic and half as non-toxic.

Afterwards, we used the following formula to measure the accuracy percentage:

$$\frac{Correct\ Classification}{Total\ Numer\ Of\ Memes}$$

Equation 1: Accuracy Method

The results were:

1. Text Classification Accuracy: 69%
2. Image Classification Accuracy: 69%
3. Meme Classification Accuracy: 72%
4. Overall Classification System Accuracy: 75%

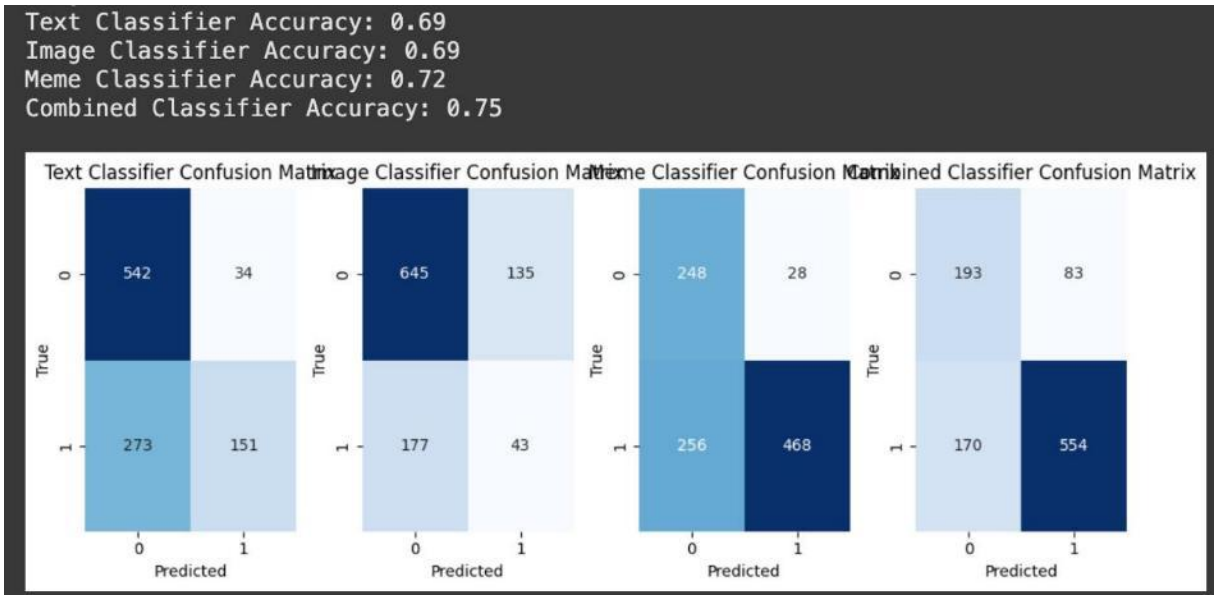


Figure 4: Classification Components Accuracy

For measuring the explanation and generation model of an alternative meme description, we used the official results of the Gemma-3 model.

Benchmark	Gemma 3 PT 4B	Gemma 3 PT 12B	Gemma 3 PT 27B
COCOcap	102	111	116
DocVQA (val)	72.8	82.3	85.6
InfoVQA (val)	44.1	54.8	59.4
MMMU (pt)	39.2	50.3	56.1
TextVQA (val)	58.9	66.5	68.6
RealWorldQA	45.5	52.2	53.9
ReMI	27.3	38.5	44.8
AI2D	63.2	75.2	79.0
ChartQA	63.6	74.7	76.3
VQAv2	63.9	71.2	72.9
BLINK	38.0	35.9	39.6
OKVQA	51.0	58.7	60.2
TallyQA	42.5	51.8	54.3
SpatialSense VQA	50.9	60.0	59.4
CountBenchQA	26.1	17.8	68.0

Figure 5: Gemma-3 Official Tests Results

For the image generation results, we used the official outputs of the Stable Diffusion model.

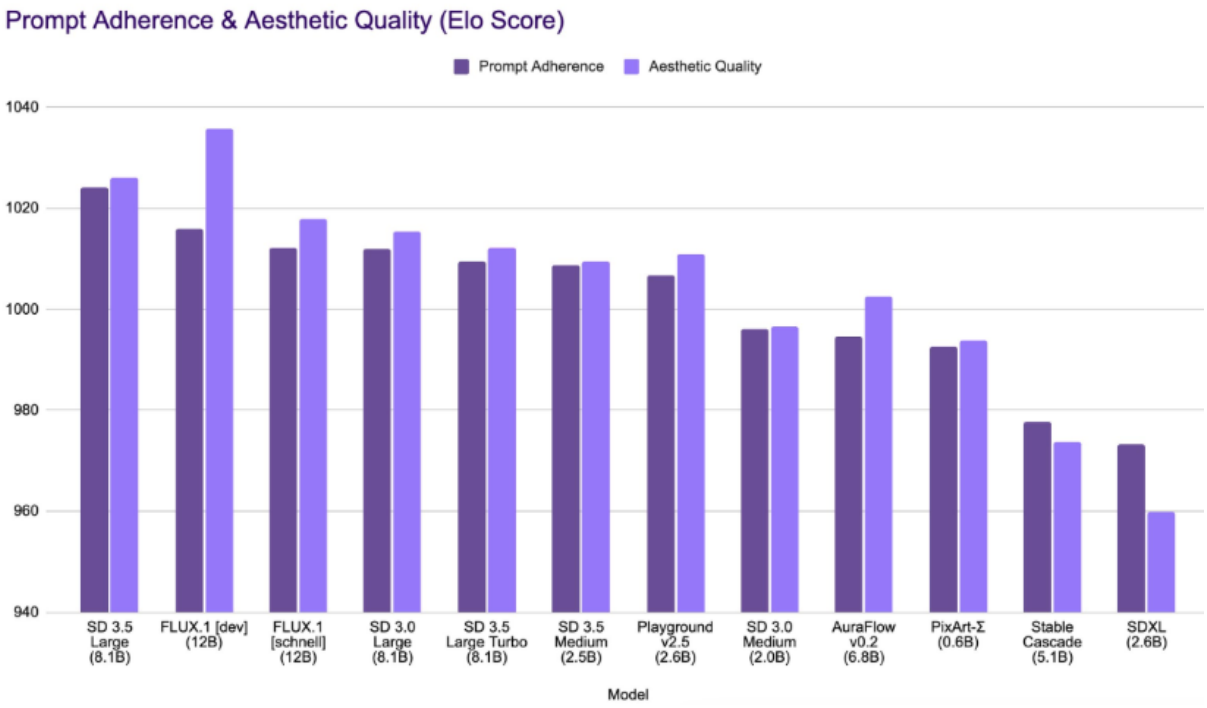
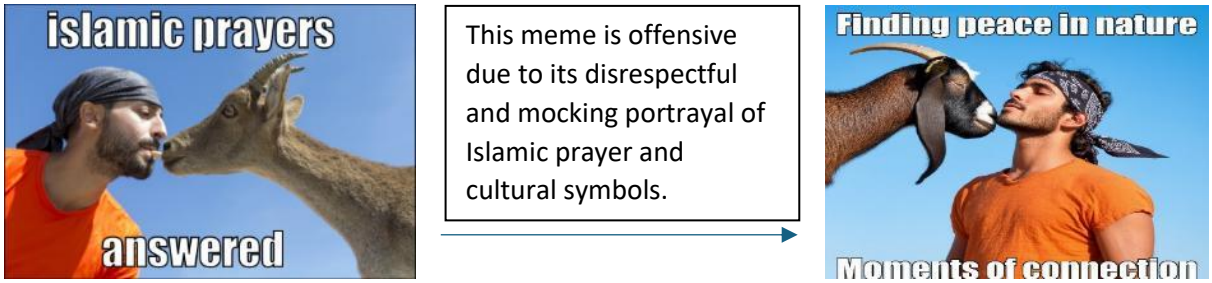


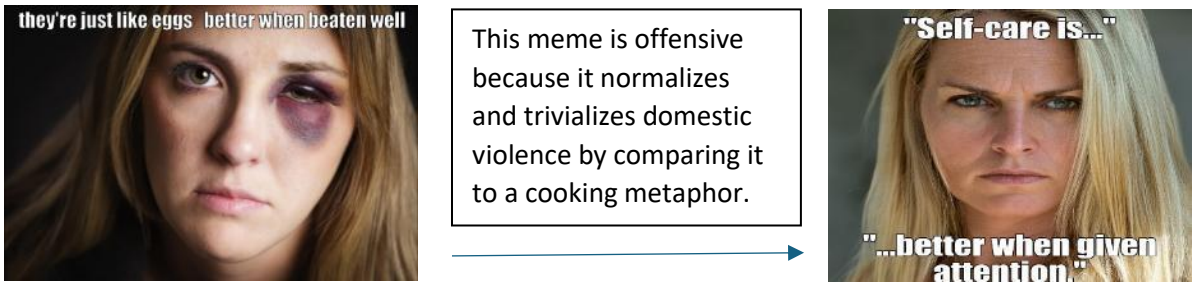
Figure 6: Stable-Diffusion 3.5 Official Tests Results

Representative Examples from the System

To demonstrate the functionality and performance of the system in real-world scenarios, we present several representative examples generated and processed during development and evaluation. These examples highlight the system's ability to handle diverse inputs. Apply accurate classification or generation logic, and produce meaningful outputs aligned with the project's objectives.

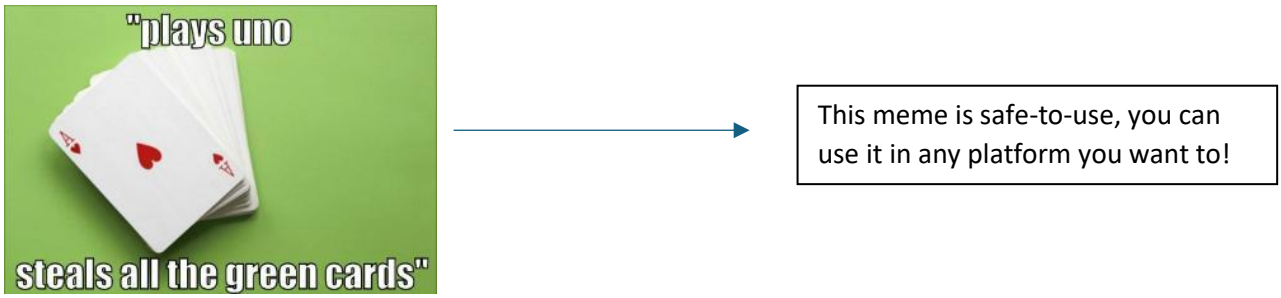


Example 1: Offensive Meme Explained and Generated using the System



Example 2: Offensive Meme Explained and Generated using the System

Example 3: Offensive Meme Explained and Generated using the System



Example 4: Non-Offensive Meme

Problems and their solutions

During the project, we encountered several challenges:

1. Lack of a sufficiently large labeled meme dataset to train a model capable of producing tailored results. The largest available dataset contains 10,000 memes. We addressed this issue by leveraging pre-trained models and dividing the system into separate tasks, each of which could be solved using existing models.
2. Limited hardware memory capacity constrained our ability to use large, powerful models that achieve the highest accuracy. We had access to 10 NVIDIA A5000 GPUs, each

equipped with 24GB of VRAM. To overcome this, we utilized smaller models and distributed the workload across multiple GPUs in a decentralized manner.

3. A dataset with many incorrectly labeled samples. To solve this, we created a reliable dataset by manually labeling 1,000 memes as either offensive or non-offensive, with approximately half classified as offensive.

Impact of Vision-Language Model Quality under Hardware and Budget Constrains

In our project, we observed that using a more advanced vision-language model significantly improved the accuracy of offensive meme detection and the relevance of generated meme alternatives. However, such models typically require stronger computation resources (e.g., high-end GPUs with large memory) and are often limited by access or licensing costs. Due to hardware and budget constraints, we were unable to fully utilize state-of-the-art models, which had a measurable effect on the system's overall performance.

Here are the same examples presented earlier, processed through the LLava-4-Scout [8] interface – a large, state-of-the-art vision-language model, and Stable-Diffusion-3.4-Large interface [7] – a large text-to-image model that we were unable to run locally due to hardware limitations.

The meme appears to mock Islamic prayers and the idea of a divine response. The image of a man and a goat kissing could be seen as disrespectful to the .Islamic faith and its practices

Example 5: offensive meme explained and generated using larger models

The meme is offensive because it trivializes domestic violence and makes a joke about physical abuse. The image of a woman with a bruised face, accompanied by the caption "they're just like eggs, better when beaten well," is a disturbing and insensitive comparison that downplays the .severity of domestic violence

Example 6: Offensive Meme Explained and Generated using larger models

The meme is offensive because it uses a gorilla to represent Black people, which is a form of dehumanization and perpetuates racist stereotypes. The image reinforces the idea that Black people are somehow less than human or more animalistic, which is a harmful and discriminatory notion. This kind of imagery has been used historically to justify racism and oppression, and its use in a meme can be hurtful and alienating to Black people. Additionally, the caption "how I see black people getting trolled on the internet" implies that the creator of the meme is observing Black people being harassed online, but the use of a gorilla image reduces the complexity of the issue to a simplistic and demeaning representation. Overall, the meme relies on racist tropes and stereotypes, making it .offensive and hurtful

Example 7: Offensive Meme Explained and Generated using larger models

The offensiveness could potentially :stem from
The casual tone towards a serious .topic
The misspelling of "Mediterranean," which might be seen as disrespectful .or mocking
The implication that the rescue of a group of people (potentially migrants) is somehow humorous or .not noteworthy

Example 8:Offensive Meme, That our System Classified as Non-offensive, Explained and Generated using larger models

These models demonstrated significantly improved understanding of visual context and textual nuance, resulting in better classification of offensive content and more relevant alternative meme suggestions. Its performance highlights the potential benefits we could have achieved with access to higher end computing resources, emphasizing how critical hardware and budget constraints were in shaping the capabilities and accuracy of our final system.

Conclusions and Recommendations

We recommend that if more powerful hardware becomes available, the strong multimodal model should also be tasked with classification to evaluate its performance. Additionally, a more advanced multimodal model would likely generate higher-quality descriptions for alternative memes, improving the overall system's effectiveness. We also recommend testing the system on a larger, well-labeled dataset to better assess its generalization capabilities and improve robustness.

The classification accuracy we achieved was approximately 75%, compared to the initial goal of over 80%. This shortfall was mainly due to hardware limitations, which constrained the complexity and size of the models we could use. With better hardware, we believe higher accuracy could be reached.

Furthermore, we aimed for the system to respond within 2 seconds, but actual response times were longer—again primarily due to hardware constraints.

The product effectively identifies offensive memes, explains the reasons for toxicity, and generates alternative memes that convey the same message in a non-offensive manner while preserving the original meme's character. However, the system's dependency on hardware limits its scalability and real-time performance. The manual curation of the dataset also highlights the challenge of obtaining high-quality labeled data.

Overall, the project met its main objectives by developing an end-to-end system for classification, explanation, and alternative meme generation. Despite some limitations in accuracy and speed, the approach demonstrates a viable solution to mitigating harmful content in memes.

We adhered to the timeline and budget we set at the start of the project. While hardware constraints limited some performance aspects, the project was completed within the planned scope and resources.

Task Management

Every Thursday, for approximately ten hours, we worked collaboratively on the project. The milestones of the project are as follows:

January: Completion of the text and image extraction model from memes.

February: Completion of the classification model, integrating the explanation model.

March: Completion of the meme generation model.

April: Integration of all models into a single system

May: Completion of the API application that incorporates the system.

May: System testing and improvement

June: Completion of the Project

References

[1] Meta, Hateful Memes Challenge, Hateful Memes Challenge and dataset, May 2020.

[2] Haoran W. Chenglong L. Jinyue C. Jia W. Lingyu K. Yanming X. Zheng G. Liang Z. Jianjian S. Yuang P. Chunrui H. Xiangyu Z., General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model, arXiv:2409.01704, Sep 2024.

[3] IBM, IBM Granite, Granite | IBM, Sep 2023.

[4] OpenAI, CLIP, CLIP/model-card.md at main · openai/CLIP · GitHub, Apr 2022.

[5] Cao Y. Wu J. Alistair C. L. C. Bryan S. G. Theodore L. C. J. Sherman C. Z. S., Detecting Offensive Memes with Social Biases in Singapore Context Using Multimodal Large Language Models, arXiv:2502.18101, Feb 2025.

[6] Google, Gemma-3, סקירה כללית של דגם Gemma 3 | Google AI for Developers, Mar 2025.

[7] Patrich E. Sumith K. Andreas B. Rahim E. Jonas M. Harry S. Yam L. Dominik L. Axel S. Frederic B. Dustin P. Tim D. Zion E. Kyle L. Alex G. Yannik M. Robin R., Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, arXiv:2403.03206, Mar 2024.

[8] Meta, Llama-4, The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, Apr 2025.

המלצת ציון לדו''ח מסכם

<u>אם יש צורך, לכל סטודנט/ית בנפרד</u>

מספר הפרויקט: _____-____P-20, שם הפרויקט:

שם המנחה:

שם הסטודנט/ית: ת.ז.:

%	קריטריון	1 - חלש	2 - בינוני	3 - טוב	4 - ט"מ	5 -מצוין
9	הגדרת המטרה - האם מטרת הפרויקט ברורה? המטרה צריכה לכלול גם את התרומה הצפויה מהשלמת הפרויקט.					
10	ארכיטקטורת הפתרון - האם הובהר מדוע הארכיטקטורה המוצעת לפתרון ההנדסי מתאימה לפתרון הבעיה? האם הייתה התייחסות לאלטרנטיבות?					
9	הצגת התוצאות - האם הוסברו המשמעויות של התוצאות, או המסקנות הנובעות מהן, האם רק הייתה הצגה עובדתית של תוצאות?					
9	מבנה – מצגת - האם יש מבנה הגיוני להצגת העבודה? מבוא, רקע ו-SoTA, מטרות, שיטה, פתרון, תוצאות, מסקנות.					
9	רמת קושי - כיצד הנך מעריך את הפרויקט בהשוואה לפרויקטים אחרים השנה או בשנים קודמות?					
9	איכות כתיבה / הצגה - נא להתייחס לאיכות הצגת העבודה: תחביר, מבנה וקשר הגיוני בין החלקים במהלך הצגת העבודה					
9	שורה תחתונה - האם מטרות הפרויקט הושגו? יש להתייחס למטרות כפי שהוגדרו בתחילה.					
9	הבנה - האם נראה שהסטודנט מבין את העבודה שנעשתה? האם עומק הדיון מספק?					

9	השקעה - האם ניכר שהסטודנט השקיע בעבודה בפרויקט ו/או בהכנת ההצגה של העבודה? האם נראה שהעבודה נעשתה בחופזה, כדי לצאת ידי חובה?					
9	דירוג - כיצד הנך מעריך את איכות הפרויקט ביחס לכלל הפרויקטים שראית?					
9	המאפיינים העיקריים בפרויקט - יש לסמן את כמות המאפיינים שהושגו (ולו באופן חלקי). מאפיין אחד בעמודה 1 וכן הלאה... המאפיינים: * מימוש/יצור מערכת, * אנליזה מתמטית, * תכנון אלגוריתמי, * אינטגרציה עם מערכת קיימת, * חקר ביצועים.					