Habituate: AI-Powered Habit Transformation*

Prateek Mohan
MCS
Arizona State University
Tempe, AZ, USA
pmohan9@asu.edu

Aditi Ganapathi
MS CS
Arizona State University
Tempe, AZ, USA
aganap12@asu.edu

Asutosh Karanam
MS CE
Arizona State University
Tempe, AZ, USA
akarana5@asu.edu

Hetvi Shah
MS DS
Arizona State University
Tempe, AZ, USA
hshah119@asu.edu

ABSTRACT

Developing and sustaining positive lifestyle habits is a common challenge, as individuals often struggle to recognize subtle cues of waning motivation or disruptions to routines that can lead to habit abandonment. Traditional habit tracking approaches relying solely on manual logging lack the nuance to detect these early warning signs. This research proposes an AI-powered, image-based habit tracking mobile app called Habituate to address this limitation.

Habituate incorporates computer vision for workplace tidiness and food image analysis to provide users with insights into their organization and dietary patterns, promoting more informed, personalized feedback. Additionally, it employs anomaly detection on workplace images to identify potential motivational roadblocks like disorganization for the habit the user is trying to cultivate. The hypothesis is that Habituate's live advising functionality will enhance self-awareness, enable early intervention for setbacks, and ultimately increase habit formation success rates compared to conventional methods.

The app's workflow involves uploading food and workspace images, which undergo object detection using ResNet. The detected objects are then processed by Moondream2 Visual language model and Mistral-7B in combination with RAG to generate tailored tips for dietary improvement and maintaining an organized workspace conducive to productivity. The study evaluates Habituate's performance using various object detection algorithms, datasets, and language models, with promising initial results suggesting its potential to revolutionize habit tracking through AI integration.

KEYWORDS

Object Detection, Habit Tracking, Generative AI, Artificial Intelligence, Application Development, Android, Django Server, Web API

*Demo video for the project can be found here.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CSE 572: Data Mining, Spring 2024, Tempe, Arizona, USA

ACM Reference format:

Prateek Mohan, Aditi Ganapathi, Asutosh Karanam, and Hetvi Shah. 2024. Habituate: AI-Powered Habit Transformation. In the Data Mining Final Project (CSE572-Spring 24). ASU, Tempe, AZ, USA, 8 pages.

1 Introduction

The lifestyle choices we make on a daily basis, such as our eating habits, exercise routines, recreational drug use, and preventative health actions, have a profound impact on our overall health risks. These risk factors are not static, but rather dynamic and modifiable based on our individual behaviors and activities. Addressing and positively influencing these modifiable lifestyle factors is critical for any effective healthcare intervention to truly improve health outcomes.

For a significant portion of the population, developing new habits (habit formation) and successfully integrating them into their daily routines over the long-term (habit maintenance) poses a considerable challenge. One key obstacle is the ease with which subtle signs of waning motivation or disruptions to established routines can go unnoticed, quickly leading to the abandonment of positive habits. Furthermore, most contemporary habit tracking applications rely heavily on manual logging by the user, which lacks the sophistication and contextual awareness to detect these early warning signals of potential habit failure.

1.1 Proposed Solution

To address these limitations, we propose Habituate - an AI-driven mobile application that leverages advanced computer vision and natural language processing techniques to provide intelligent, personalized habit tracking and support. By allowing users to log habits through uploaded images rather than just manual text entry, Habituate can analyze visual data for deeper insights into the user's behaviors and patterns. The key features of Habituate include:

1.1.1 Image Analysis. Image analysis for dietary habit tracking by detecting foods and providing tailored nutritional feedback.

1.1.2 Anomaly Detection. Anomaly detection of disorganization/clutter in workplace images as an early cue of decreasing productivity.

1.1.3 Tip Generation. Integration of large language models to generate motivational tips and accountability prompts based on the user's uploaded images and interaction history.

It is hypothesized that Habituate's multimodal approach, combining image logging with AI-powered analysis and tailored language outputs, will lead to increased self-awareness, earlier intervention for potential setbacks, and ultimately higher rates of successful long-term habit formation and maintenance compared to traditional manual tracking methods.

1.2 Object Detection

At a high level, object detection models take an input image and output a set of bounding boxes, each with an associated class label and confidence score. This allows the model to not only identify what objects are present, but also where they are located in the image. Deep convolutional neural networks (CNNs) are used as the foundation for feature extraction in many contemporary object detection algorithms. These CNN backbones, which include ResNet, VGG, and EfficientNet, are essential to the overall object detection system's functionality.

1.2.1 ResNet (Residual Neural Networks). ResNet (Residual Neural Network) is a CNN architecture that introduces residual blocks with skip connections to address the degradation problem in very deep neural networks. These residual blocks allow the input to be added directly to the output of stacked convolutional layers, creating a bypass path. This innovative design enables efficient gradient flow during training, even in extremely deep networks with hundreds or thousands of layers. By mitigating the vanishing gradient issue, ResNet revolutionized the depth capabilities of CNNs, paving the way for substantial performance improvements on challenging computer vision tasks like that of Habituate’s.

1.2.2 YOLOv5 (You Only Look Once v5). YOLOv5 is a state-of-the-art object detection algorithm that employs an anchor-based, one-stage approach to simplify the detection process. It utilizes a pre-trained backbone network for feature extraction and then predicts bounding boxes and object class probabilities in a single forward pass. This efficient design, coupled with optimizations, enables YOLOv5 to achieve real-time object detection with high accuracy across various applications. It strikes an impressive balance between speed and performance, making it the choice for tasks requiring fast and reliable object localization and recognition fitting the exact use case of our research methodology.

1.3 Vision Language Models (VLMs)

Vision-language models (VLMs) are a category of artificial intelligence models that can comprehend and produce textual and visual material by combining computer vision with natural language processing skills. These models have become an effective means of bridging the gap between language and visual intelligence, opening a plethora of applications in domains such as virtual and augmented reality, accessibility, content creation, and healthcare. Generally, VLMs use one of two primary architectural approaches: either a single encoder that can handle both modalities, or multimodal fusion, which combines independent encoders for textual and visual inputs.

1.3.1 Moondream2. Moondream2 is a state-of-the-art vision-language model. With independent encoders for textual and visual inputs and a pivotal multilayer perceptron projector that connects the two representations, the model uses a multimodal fusion technique. This architectural layout and the model's 1.86 billion parameters, which are relatively light, enable effective deployment on edge devices without sacrificing performance on a wide range of vision-language tasks. The Moondream2 architecture differs from larger, more resource-intensive VLMs in that it prioritizes efficiency and edge device deployment. Moondream2 intends to expand the scope of vision-language models' capabilities to a broader variety of devices and applications by utilizing a more compact and effective design.

1.3.2 Moondream2 Architecture. Moondream2 employs a multimodal fusion approach, with separate encoders for visual and textual inputs. The visual encoder is based on the SigLIP-400M model, while the language encoder is derived from the Phi-1.5 backbone. A crucial addition to the Moondream2 architecture is the incorporation of an MLP (multilayer perceptron) projector. This component bridges the visual and textual representations, potentially enhancing the model's vision-language alignment and improving its performance across various tasks.

1.3.3 Moondream2 Pre-training. A wide range of datasets were used to train Moondream2, including 220,000 photos from the LVIS-INSTRUCT4V dataset, 60,000 photos from the ShareGPT4V dataset, 150,000 calls to private functions, and 50,000 exchanges from the OpenHermes-2.5 dataset. This extensive training set guarantees Moondream2's ability to perform well on a variety of vision-language tasks, including conversational understanding, object identification, and instruction following.

1.4 Large Language Models (LLMs)

Large language models (LLMs) use the transformer architecture to accomplish a variety of natural language processing tasks. The transformer processes the input sequence in parallel by capturing contextual relationships among words with the use of attention mechanisms. The fundamental components are feedforward networks, which analyze the output of the attention layers, and attention processes, which assess the significance of each word. The transformer stack's depth makes it possible for LLMs to pick up intricate linguistic patterns.

Large text corpora are optimized for language modeling during the training phase, which helps the models gain a thorough grasp of language structure, semantics, and context. One important reason for LLMs' remarkable success on a variety of tasks is their enormous scale—they have hundreds of billions of parameters.

1.4.1 MISTRAL Mixtral 7B. Mixtral 7B is a LLM developed by the French AI startup Mistral. It makes use of a cutting-edge attention mechanism known as Grouped-Query Attention, which splits the query heads into groups, each of which has a single key and value head. This offers better performance and efficiency as an interpolation between multi-head and multi-query attention. It uses a sliding window attention method that keeps its cache size

fixed and gives it a long effective attention span (up to 128K tokens). Character mapping to tokens that are not in the vocabulary is prevented by using the Byte-Pair Encoding tokenizer.

1.4.2 Retrieval Augmented Generation (RAG). Retrieval-Augmented Generation is a strategy that enhances the quality and relevance of generated responses by fusing the capacity of LLMs with the information contained in external data sources. A robust RAG pipeline for developing sophisticated conversational AI applications is produced by combining LangChain, Weaviate, and Mixtral-7B. A flexible framework called LangChain makes it possible to incorporate the powerful natural language processing capabilities of the massive language model Mixtral-7B into intricate workflows. The scalable vector store Weaviate functions as the external knowledge base that the RAG model may query to obtain pertinent data and enhance the results produced by Mixtral-7B. Through this integration, the system may make use of a larger body of information that goes beyond the training data for the model, leading to more intelligent, pertinent, and interesting conversations. The RAG pipeline constructed with LangChain, Weaviate, and Mixtral-7B is an effective approach for creating sophisticated conversational AI systems that can smoothly integrate language understanding, creation, and information retrieval by utilizing the advantages of each component.

2 Related Work

While existing work has explored habit formation, behavior change, and AI applications in fields like nutrition and workspace organization, there are limitations in effectively addressing the challenges of manual logging and lack of sustained motivation during habit formation. The individual actions of people, whether related to eating, exercise, recreational drug use, or preventive and rehabilitative activities, directly impact the health risks associated with contemporary lifestyle. These risks must be modified for users to be effectively treated [1]. Lally & Gardner [2013] and Rebar et al. [2022] provide theoretical foundations and empirical evidence for understanding habit formation mechanisms and effective intervention strategies [2, 3]. However, they do not directly address the specific issues of manual logging and maintaining motivation over time.

The works on transformer technology in object detection, such as Carion et al. [2020] and Zhu et al. [2020], demonstrate the potential of transformers for simplifying object detection pipelines and improving performance, especially for small objects. While these advancements could be leveraged for image-based logging and anomaly detection in the proposed AI-powered habit tracker, they do not directly tackle the motivational aspect of habit formation [4, 5]. Similarly, studies on AI in nutrition and workspace organization, like Zhang & Wallace [2018], Ciocca et al. [2017], and Afzal et al. [2023], offer insights into using CNNs and transformers for feature extraction and object detection from images [6-8]. However, they primarily focus on the technical aspects of image analysis rather than integrating these capabilities into a comprehensive system for habit formation and sustained motivation. The ML-based system proposed by Gowthamani et al. [2022] analyzes user activity data obtained via a web application using machine learning methods. Relevant information is obtained from the user at login. After that, a binary classification algorithm trained on earlier user datasets is applied to this data. Whether a user's behavior is normal or needs care is determined by the system. With tasks completed, users can keep track of their progress and export the data as a PDF [9]. Habit Driven, an AI-powered app, assists in the development of habits and goal-achieving. It provides consumers with an all-encompassing method for encouraging beneficial behavioral changes through the use of tailored tracking, professional advice, and informative material delivery. The user logs their daily activity in a form, which is analyzed using AI for goal tracking, based on which a chatbot provides insights upon request [10].

The existing works cited in the literature review, while valuable in their respective domains, do not directly address the specific challenges of manual logging and lack of sustained motivation that can hinder habit formation efforts.

3 Methodology

This section elucidates the methodology adopted to complete this project. It involves 7 phases, of which 3 are pursued concurrently due to their related nature as well as to optimize the utilization of time. Figure 1 details these phases.

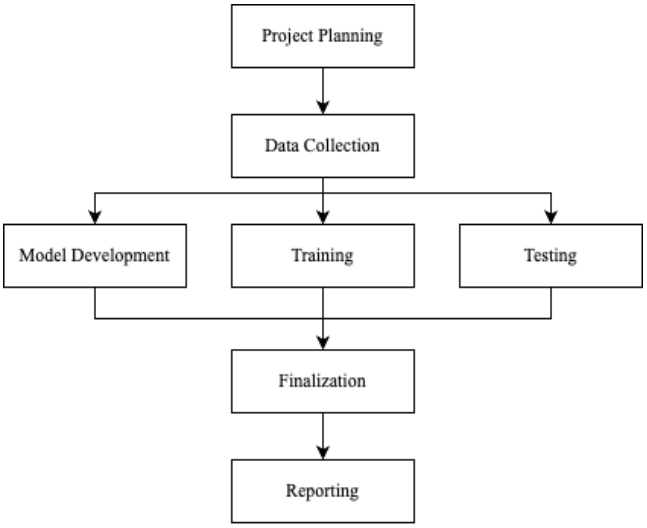


Figure 1: Adopted methodology

3.1 Project Planning

This phase involved deciding the flow of the application and its implications in the backend. Figure 2 details the result of this phase.

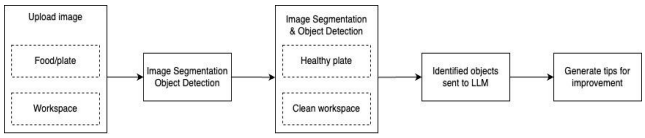


Figure 2: Application workflow

The user uploads an image of either their food plate or their workspace, which will trigger image segmentation and object detection in the backend to identify the contents of their plate or the objects in their workspace. With the identified objects, the information is sent to an LLM, which will generate tips to improve the contents of the plate or the state of the workspace. The models for object detection are also selected in this phase.

Due to the computationally intensive nature of this project, it was planned that Google Colaboratory's T4 GPUs and ASU Sol's A30 GPUs with at least 8 CPU cores would be employed for training the models.

3.2 Data Collection

In this phase, the datasets required for training the selected models were chosen after thorough study of various publicly available datasets. For food detection, the BeFit dataset was chosen from Roboflow with 36 classes of various dishes, containing 2,453 training, 703 validation, and 351 testing images [11]. For workspace cleanliness, a diverse set of synthetic images depicting various workstation scenarios were collected to train and evaluate the object detection models. The Synthetic Image Image Dataset v4 from Roboflow was chosen, which consists of 2,300 images split into 2,000 training images (95%), 200 validation images (3%), and 100 test images (2%). The dataset underwent preprocessing steps such as auto-orientation and resizing to a consistent resolution of 640x640 pixels [12].

The comprehensive synthetic workstation dataset and the BeFit food dataset will allow the object detection models to be trained on diverse scenarios, ensuring accurate detection and classification of objects in real-world settings.

3.3 Model Development, Training, and Testing

In this phase, the tasks of developing the models, training them on the selected datasets, and evaluating their performance were conducted. For object detection, several state-of-the-art models were tested, including ResNet, RetinaNet, Visual Transformers, YOLOv5, and YOLOv8. These models were initialized with pre-trained weights and fine-tuned on the selected datasets to adapt them to the specific task of detecting objects in workstation and food images. The training process involved optimizing the model parameters using the training data and monitoring the model's performance on the validation set. The training loss, validation loss, and other relevant metrics such as bounding box loss (loss_bbox) and cross-entropy loss (loss_ce) were tracked to assess the models' learning progress and convergence. These metrics provide insights into how well the models are learning to localize objects (loss_bbox) and classify them correctly (loss_ce). After training, the models were evaluated on a separate test set to assess their generalization performance and compare their effectiveness in detecting objects accurately. The evaluation metrics used include precision, recall, and F1 score, which measure the models' ability to correctly identify objects while minimizing false positives and false negatives.

To further improve the models' performance and robustness, additional training iterations will be conducted over time. This will involve incorporating more diverse training data, fine-tuning hyperparameters, and exploring techniques such as data augmentation and transfer learning. The goal is to continuously enhance the models' ability to detect a wider range of objects and adapt to various real-world scenarios.

Since the app also has a text generation component apart from the object detection component, this phase also included identifying the most performant VLM and LLM for generating personalized tips and recommendations based on the detected objects. This involved training the LLMs on relevant domain-specific data and optimizing them for generating coherent and actionable suggestions. All metadata related with the image, including detected object classes, bounding box positions, and previous conversation history, were stored in vector databases in the form of text files to provide the LLM with additional context.

3.4 Finalization

The finalization phase involved integrating the selected pre-trained models to create a cohesive and user-friendly application. After evaluating the performance of various object detection models, including ResNet, RetinaNet, Visual Transformers, YOLOv8, and YOLOv5 on a set of workstation and food images, ResNet was finalized for the workspace cleanliness task, while YOLOv5 was selected for the food assessment task.

Integrating the chosen pre-trained models into the application pipeline ensured smooth interaction between the object detection component and the other modules, such as image preprocessing and tip generation. Attention was given to optimizing the model's performance, fine-tuning it if necessary, and creating a seamless user experience.

3.5 Reporting

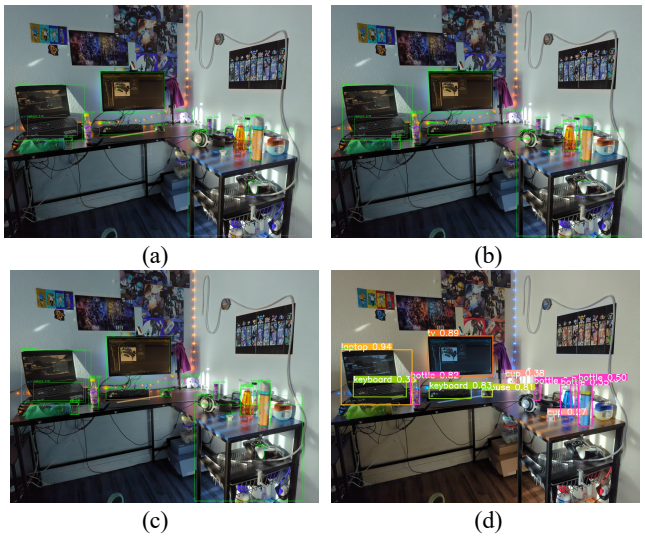
The reporting phase involved preparing a comprehensive report that details the entire development process, including the implementation details, analysis, and discussion of the results obtained at the end of the project. The report covered the selection of datasets, such as the BeFit dataset for food detection and the synthetic workstation dataset for workspace cleanliness assessment, as well as the training and evaluation of various object detection models, including ResNet, RetinaNet, Visual Transformers, and YOLOv8.

4 Results

The purpose of this section is to thoroughly assess the capabilities and performance of the suggested Habituate system architecture. Through the utilization of a range of innovative techniques such as anomaly detection, computer vision, and advanced language models, our goal was to extend the limits of what is conceivable in the field of habit tracking and formation.

4.1 Workspace Organization

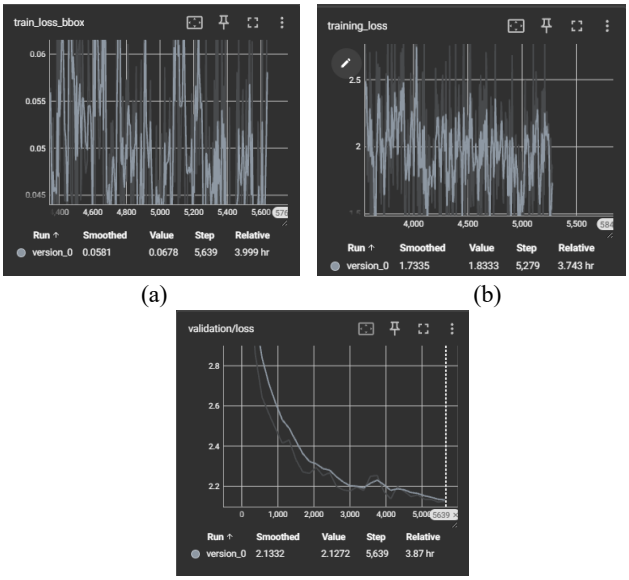
The evaluation of the object detection models—ResNet, RetinaNet, Visual Transformers, and YOLOv8—was conducted using a set of workstation images. Figures 3 (a), (b), (c), and (d) discuss the qualitative differences between the models.



(a) (b) (c) (d)

Figure 3: Object detection results with (a) ResNet, (b) RetinaNet, (c) ViT, and (d) YOLOv8

ResNet had a strong ability to detect larger objects, such as monitors and chairs, with high confidence; however, it occasionally missed smaller items like pens or notebooks. RetinaNet performed well in detecting both large and small objects. While its precision in bounding box placement was notable, it suffered from multiple bounding boxes. ViT excelled in recognizing objects with complex shapes and in cluttered scenes. It was particularly good at identifying overlapping objects. YOLOv8 demonstrated exceptional speed and accuracy. It reliably detected a wide range of objects with high confidence. The bounding box training loss, training loss, and validation loss have been encapsulated in Figures 4 (a), (b), and (c).



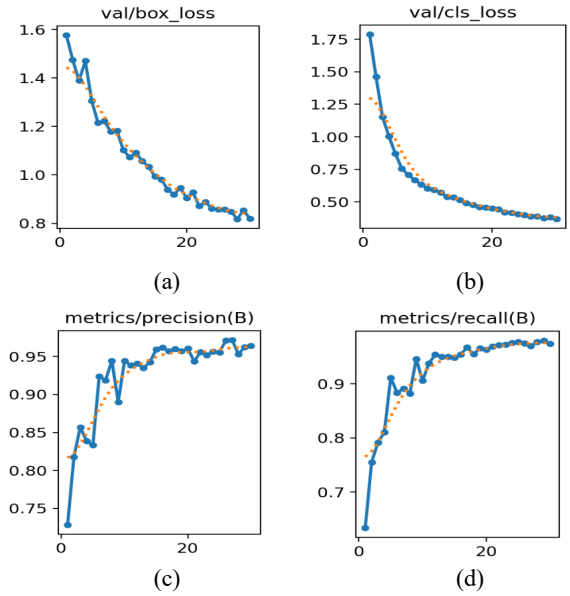
(a) (b) (c)

Figure 4: Training and validation losses

ResNet performed as well as YOLOv8, but with the aim to explore diverse models for the app, ResNet was selected due to its transformer-based architecture.

4.2 Food Analysis

It was observed that most models were unable to capture the nuances in the dish presented on the plate, and often mislabeled similar-looking dishes – this resulted in a poor performance in terms of accuracy. However, YOLOv5 presented the most promising results with a precision and recall of 95%. Figure 5 (a), (b), (c), and (d) show the validation box loss, class loss, precision, and recall.



(a) (b) (c) (d)

Figure 5: The (a) box loss, (b) class loss, (c) precision, and (d) recall observed with YOLOv5

The model was trained for 30 epochs and saw an excellent decrease in box and class loss, with increasing precision and recall. However, a limitation was the number of classes of dishes that could be detected by the model. Figure 6 shows the object segmentation for Chinese takeout with multiple small components, all identified correctly by the model.

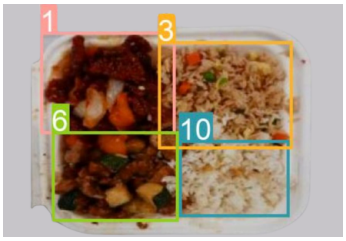


Figure 6: Accurate bounding box allocation

Despite being trained on a 36-class dataset, real-world situations often encounter dishes from various cuisines and cultures, requiring a much more diverse dataset to train the model. For the limited dataset provided, YOLOv5 performed well.

4.3 VLM, LLM, and RAG

Moondream2 performed extremely well in detecting specific objects in the workspace, while it could only generally detect food items on a plate. It had an impressive capability of answering questions with direct context from the image but was not very robust in answering abstract questions or generating tips.

Mistral, a heavier model with 7B parameters, performed much better in terms of generating tips, given context from the image as well as previous conversation history. A variety of different open-source vector databases were tested for indexing speed, including FAISS, Weaviate, and ChromaDB. FAISS had a higher indexing time compared to Weaviate, while ChromaDB was more complicated to use than Weaviate. Essentially, Weaviate was the most performant vector database for Habituate.

5 Conclusion

The findings of this study show how the Habituate app, which integrates cutting-edge AI technologies, has the exciting potential to transform habit tracking and development. The efficiency of the system has been reinforced by several important discoveries from the conducted study.

Results show that ResNet did remarkably well in the object detection domain in terms of correctly classifying workstation cleanliness, which is essential for sustaining productivity and forming healthy behaviors. On the other hand, YOLOv5 turned out to be the best option for food image analysis, which allowed Habituate to give consumers individualized insights into their eating habits. Based on the identified objects and abnormalities, the Mistral-7B model-powered language modeling component of the system demonstrated remarkable efficacy in producing pertinent and customized feedback for users. Moreover, the system's capacity to offer contextually-rich and educational suggestions to assist habit formation was further improved by the incorporation of the RAG technique, which makes use of the Weaviate vector database.

These results highlight the benefits of Habituate's AI-powered habit monitoring methodology, which surpasses the constraints of conventional human logging techniques. Through the utilization of computer vision, anomaly detection, and cutting-edge language models, Habituate enables users to gain a more profound comprehension of their routines and obtain prompt interventions to surmount possible obstacles.

5 Future Scope

The encouraging findings of the study have prepared the way for Habituate's next stage, which will involve creating a complete mobile application supported by a strong Django server architecture. With this upgrade, the system will be able to offer users individualized feedback and real-time habit tracking, which will significantly improve its ability to promote long-term lifestyle modifications.

The diversification of the food detection training dataset is a major area of attention for future growth. The system will be able to identify a greater range of food items and provide users with more thorough dietary insights if the breadth and depth of the food images used to train the computer vision models are increased. There is also an intention to use LLMs that have been specifically trained on dietary guidelines and nutritional advice to further enhance the tip generating portion of food identification. With the use of this unique language modeling technique, Habituate will be able to offer consumers even more customized, relevant, and useful advice to help them maintain healthy eating practices.

ACKNOWLEDGMENTS

The authors would like to sincerely thank Professor Hua Wei for his excellent advice and knowledge during the Habituate project's development. We also sincerely appreciate Hao Mei, the teaching assistant, for his unceasing efforts in offering timely and constructive feedback that has been crucial in improving the technical parts of our work. The authors also thank the course graders for their invaluable contributions, whose in-depth assessments have strengthened the Habituate's general quality and helped us discover areas for development.

REFERENCES

[1] Sarbadhikari, Suptendra Nath, and Jyotika Maggo Sood. "Gamification for nurturing healthy habits." The National Medical Journal of India 31, no. 4 (2018): 253-254.

[2] Lally, P., & Gardner, B. (2013). Promoting habit formation. Health Psychology Review, 7(sup1), S137-S158. DOI:https://doi.org/10.1145/567752.567774

[3] Rebar, A. L., Dimmock, J. A., Jackson, B., Rhodes, R. E., Kates, A., Starling, J., & Vandelanotte, C. (2022). A systematic literature review of the effects of priming on habit formation. Health Psychology Review, 16(1), 1-26.
[4] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In European Conference on Computer Vision (ECCV).
[5] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.
[6] Zhang, Y., & Wallace, B. C. (2018). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. Journal of Machine Learning Research, 19(1), 1-37.
[7] Ciocca, G., Napoletano, P., & Schettini, R. (2017). Food recognition: a new dataset, experiments, and results. IEEE Journal of Biomedical and Health Informatics, 21(3), 588-598.
[8] Afzal, M. Z., Mahmood, A., Hussain, M., & Hayat, K. (2023). Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art. arXiv preprint arXiv:2309.04902.
[9] Gowthamani, R., K. Sasi Kala Rani, M. Indira Priyadharshini, M. Rohini, Grace Ebenezer, and Emma Thomas. "Web Based Application for Healthy Habit Development Through Gamification with ML." In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1338-1345. IEEE, 2022.
[10] Mike Jalonen. 2023. It's time to be habit driven…. Habit Driven. https://habitdriven.ai/
[11] BeFit. 2023. BeFit Dataset. Open-Source Dataset. Roboflow Universe. Roboflow. https://universe.roboflow.com/befit/befit-xjs26 (visited on 2024-04-19).
[12] University of Bath. 2023. Synthetic Image Dataset. Open Source Dataset. Roboflow Universe. Roboflow. https://universe.roboflow.com/university-of-bath-mjtna/synthetic-image (visited on 2024-04-19).