AI-Driven Usability Testing: Integrating Eye-Tracking Data and Agentic Systems for Automated UI Evaluation

Mehweesh Kadegaonkar, Kayvan Karim

Heriot-Watt University
mehweeshtk@gmail.com, K.Karim@hw.ac.uk

Abstract

Despite the benefits of user interface/experience (UI/UX) design, traditional usability testing remains resource-intensive and repetitive. This study proposes a novel system that integrates real-time browser-based eye-tracking with a multimodal agentic framework to automate UI evaluation. Participants interacted with task-specific interfaces while their gaze data was captured and analyzed by a multi-agent system to generate structured usability reports grounded in heuristic principles. Precision metrics were used to quantify qualitative insights, enabling measurable evaluation. To enhance accessibility, a comparative analysis was conducted between proprietary and open-source Large Language Models (LLMs). Results showed that proprietary models consistently delivered accurate insights, whereas smaller local models struggled with reliability — highlighting future directions for offline deployment. The findings contribute to the advancement of AI-driven solutions in usability evaluation, showcasing how agentic systems integrated with browser-based eye-tracking tools can overcome traditional limitations.

Introduction

The quality of user interface and experience (UI/UX) design influences our digital experience profoundly, often going unnoticed until it's done wrong. Effective UI/UX enhances visual appeal, communicates brand identity, and ensures intuitive user experiences. With the continuous evolution of Artificial Intelligence (AI), the field of UI/UX has also been evolving (Bertão and Joo 2021). Researchers are delving into the application of AI tools in five key areas of the design process: understanding the context of use, user requirements, solution design, evaluating design, and development of solutions (Stige et al. 2023).

However, design evaluation remains a particularly repetitive and time-consuming step, requiring multiple rounds of refinement to meet user needs (Stige et al. 2023). Eye-tracking, although long explored in usability research (Jacob 2002; Poole and Ball 2004), has faced adoption barriers due to its reliance on expensive, specialized equipment.

Recent studies have introduced AI-driven tools for UI generation, such as MetaMorph (Pandian et al. 2020), which uses computer vision to convert sketches into high-fidelity prototypes, and Paper2Wire (Buschek, Anlauff, and Lachner 2020), which applies machine learning to digitize hand-drawn wireframes. Despite these advancements, limited research has been conducted on the integration of agentic systems — a recent advancement in AI — into the UI/UX design process. To the best of our knowledge, no studies have utilized agentic systems within the field of UI/UX. This study takes an innovative approach by integrating multimodal agentic systems and real-time eye-tracking technology for UI evaluation.

This study aims to improve the efficiency and accessibility of usability testing, particularly in developing regions where internet connectivity is limited or costly. By automating UI evaluation and minimizing repetitive tasks, the proposed system enables UI/UX designers to focus on higher-level creative decisions and design refinement. Eye-tracking data provides an additional dimension of input that is otherwise challenging to obtain, offering unique insights into user behavior. Additionally, the study also quantifies the resulting qualitative usability data by utilizing precision metrics.

To support broader accessibility, the system was evaluated using both state-of-the-art Large Language Models (LLMs) and smaller, local open-source models. While proprietary models produced more accurate and structured results, open-source alternatives were less reliable. These findings highlight the need to refine the system architecture for local deployment, enabling fully offline processing of eye-tracking data and usability reporting — making the solution more cost-effective and accessible. There is a general consensus that AI should not be used for automating the entire process, but rather it should offer designers the tools that can make the design process easier and more efficient (Gardey et al. 2022).

The key contributions of this work are as follows:

- A webcam-based platform for real-time UI evaluation, combining gaze tracking with agentic analysis.
- Validation of the system's capability to detect usability issues and generate structured, actionable reports.
- A comparative analysis of proprietary and open-source multimodal LLMs integrated into the agentic evaluation pipeline.

- Quantification of qualitative usability feedback using precision metrics, providing a measurable assessment of system performance.

Background

Augmented AI Tools in UI Evaluation

Human feedback — such as from user studies and expert evaluations — has long been essential for refining UIs. Traditionally, this feedback relied on heuristic evaluations using predefined guidelines (Nielsen 1994; Shneiderman 1998). The field later shifted toward scalable automated assessments with tools like ARNAULD (Gajos and Weld 2005), though early efforts were limited by developer-centric evaluations. Subsequent work by (Miniukovich and Angeli 2015) expanded this direction with psychology-based metrics, which proved effective primarily for websites. Recent advances like (Haddad et al. 2024) combined eye-tracking, facial expressions, and EEG data to classify GUI designs as 'good' or 'bad.' However, their method faces three key limitations: (1) reliance on separate computational models for each modality, (2) an oversimplified binary classification that ignores nuanced design variations, and (3) dependence on expensive, impractical hardware. Parallel work by (Duan et al. 2024) investigated LLM-based heuristic feedback through a Figma plugin, but encountered low precision due to the generation of excessive irrelevant suggestions. Notably, their findings highlight the potential of multimodal LLMs to overcome these challenges — a direction that aligns with this study's goals of developing a more unified and accessible UI evaluation framework.

Multimodal LLMs

The real-world is a multimodal (MM) environment where information is perceived and exchanged in various modes, such as vision, sound, language, and touch. Augmenting LLMs with MM capabilities has witnessed significant progress, as proven by the works of (Sun et al. 2023) and (Ge et al. 2023). This evolution began with traditional cross-modal models that could only generate one modality from another (e.g., text-to-image (Rombach et al. 2021) or text-to-audio (Huang et al. 2023)) through multi-step processes. A major breakthrough came with Composable Diffusion (CoDi) (Tang et al. 2023b), which became the first model capable of simultaneously processing and generating different modality combinations. However, CoDi's lack of a core LLM limited its reasoning capabilities (Zhan et al. 2024), leading to the development of CoDi-2 (Tang et al. 2023a) with enhanced in-context learning and multimodal chat features. The field advanced further with NExT-GPT (Wu et al. 2023b), which connected LLMs with multimodal adapters and diffusion models to achieve strong performance across text-to-image, audio, and particularly video generation. While NExT-GPT demonstrated the potential of unified architectures, it faced limitations from using frozen pretrained components, resulting in alignment inconsistencies (Zhan et al. 2024). These challenges were effectively addressed by AnyGPT (Zhan et al. 2024), which introduced discrete tokenization and a novel any-to-any instruction dataset (AnyInstruct-108k) to achieve robust multimodal unification without modifying the core LLM architecture.

LLM-based Agentic Systems

In recent years, LLMs such as ChatGPT, based on the Generative Pre-trained Transformer (GPT) architecture, have gained widespread attention for their impressive performance on diverse language tasks. Researchers have increasingly leveraged LLMs to develop intelligent agents capable of addressing complex, real-world scenarios — from software development (Hong et al. 2023) to human behavior simulation (Park et al. 2023). A significant development in this space is MetaGPT (Hong et al. 2023), which implements a multi-agent system (MAS) framework using Standardized Operating Procedures (SOPs) to coordinate specialized agents (e.g., Project Managers, Engineers). While demonstrating strong performance on programming benchmarks, its limitations in handling domain-specific tasks — particularly in UI contexts — suggest the need for multimodal enhancements. The effectiveness of such MAS frameworks stems from their ability to combine specialized agents through structured collaboration (Dohan et al. 2022), which reduces hallucinations and improves reasoning (Bang et al. 2023; Talebirad and Nadiri 2023). (Talebirad and Nadiri 2023) advanced this approach through Intelligent Generative Agents (IGAs) with dynamic role allocation and feedback mechanisms, though the lack of evaluation metrics raises scalability questions. Similarly, (Shen et al. 2024)'s *Data Director* system demonstrated how pipeline-structured agent teams (analysts, designers) could generate animated data videos. While effective, its authors highlight the need for multimodal LLMs to improve performance in handling diverse input types. This study infers that while agentic systems have been successfully explored in various contexts (Li et al. 2023; Chen et al. 2023; Shen et al. 2024), they have not been fully leveraged in the field of UI/UX design.

Eye-Tracking in Usability

Eye-tracking has evolved from its origins in reading studies (Poole and Ball 2004) to become a cornerstone of UI evaluation, enabling objective assessment of element placement and visibility through gaze patterns and Areas of Interest (AOIs) (Hasse and Bruder 2015). While traditional approaches using specialized hardware (e.g., Tobii systems) provide high precision (van den Berg, Engelsma, and Peute 2024; Wang et al. 2019), their cost and accessibility limitations have hindered widespread adoption (Zelinskyi and Boyko 2024). This challenge spurred the development of accessible alternatives. TurkerGaze (Xu et al. 2015) pioneered webcam-based tracking through Amazon Mechanical Turk, using facial landmarks with Ridge Regression (RR), later refined with Support Vector Regression (SVR), to achieve scalable saliency prediction. The field advanced further with WebGazer (Papoutsaki et al. 2016), which introduced the first browser-based solution requiring no explicit calibration. It assumes gaze aligns with interaction points (e.g.,

cursor clicks) and continuously adapts using RR on 120-dimensional eye feature vectors, while operating entirely client-side. In a comparative study with the commercial Tobii EyeX, it achieved similar mean errors, validating its effectiveness. Recent analyses confirm webcams have become the dominant hardware for usability studies (Novák et al. 2023), demonstrating the field's shift toward scalable, cost-effective solutions.

Methodology

System Architecture

The proposed system automates usability evaluation by combining browser-based eye-tracking, real-time heatmap generation, and a multi-agent LLM analysis pipeline (illustrated in Figure 1). The input layer consists of a web-based interface built with HTML, CSS, and JavaScript, integrating WebGazer.js (Papoutsaki et al. 2016) for gaze capture and heatmap.js for real-time visualization as users interact with static UI images. The interface includes three key screens: a Start Screen to initiate calibration, a Calibration Screen using a 4×4 clickable dot grid with color-coded feedback (red → yellow → green), and a Homepage where users can observe the UI images. Gaze data is sent to a Flask backend, where OpenCV overlays the heatmap onto its corresponding UI screenshot. This combined image serves as a behavioral input for further analysis. The initial image and heatmap data are first processed externally via an LLM API, where the model extracts structured usability insights based on visual attention patterns — connecting user behavior directly to heuristic-based usability evaluation. These preliminary findings are then passed into the MAS - implemented using CrewAI — for further reasoning. The output layer compiles a comprehensive Markdown usability report detailing strengths, weaknesses, WCAG (W3C 2023) compliance, and prioritized recommendations. Structured output is ensured using Pydantic, maintaining consistency across agent responses.

Agentic System

Framework Selection Rationale Several agentic frameworks were considered for this study, including AutoGen (Wu et al. 2023a), LangChain, and CrewAI, each offering distinct advantages in multi-agent collaboration. AutoGen excels in dynamic conversational agents with flexible role assignments, while LangChain provides extensive modularity for integrating diverse tools and retrieval-augmented generation (RAG). However, CrewAI was selected for its structured approach to agent specialization, explicit task delegation, and seamless support for sequential workflows — critical features for the proposed system. Unlike AutoGen, which prioritizes conversational adaptability over strict task sequencing, or LangChain, which requires extensive customization for multi-agent collaboration, CrewAI enforces a clear hierarchy of roles and responsibilities. This ensures deterministic execution of heuristic evaluations and report generation, aligning with the study's need for reproducible, standards-compliant outputs. Furthermore, CrewAI's native integration with validation models (e.g., Pydantic) guarantees structured outputs, ensuring consistent, reproducible results across all evaluations.

Agent and Task Initialization The first agent, the UI/UX Recommendation Specialist (`ui_recommender`), was tasked with generating actionable recommendations by analyzing user behavior data against established usability heuristics and accessibility standards. Prompt engineering was used to enforce structured outputs, including detailed issue descriptions, heatmap correlations, and prioritized fixes. The second agent, the Report Compiler (`report_compiler`), transformed the validated recommendations into a structured Markdown report. The workflow followed a strict sequential order, with outputs from the UI Recommender feeding directly into the Report Compiler. This analysis → validation → synthesis pipeline, combined with YAML-based task definitions and prompt constraints, ensured consistent, reproducible, and standards-aligned outputs.

The original methodology initially consisted of four agents within the MAS:

- Preprocessing agent
- Analysis agent
- Recommendation agent
- Reporting agent

However, preliminary testing revealed that the initial assumption regarding the Preprocessing Agent was incorrect, as direct Python-based data preprocessing proved more efficient. An Analysis Agent was also designed to process base64-encoded screenshots, but testing showed that the model's context window could not accommodate the combined volume of image data and supplementary metadata passed through CrewAI, leading to systemic failures. To resolve this, preliminary image analysis was offloaded to a dedicated API-based LLM service, which extracted structured textual insights before forwarding them to the UI Recommender Agent for further processing. Based on these findings, the implemented MAS architecture comprises two specialized agents highlighted in Table 1.

Agent	Description
ui_recommender	Conducts heuristic usability assessments against relevant UI/UX standards, using the initial textual analysis from the API as input.
reporter_compiler	Generates usability evaluation report in markdown format, enforcing structured outputs through Pydantic validation based on the results from the Recommendation agent.

Table 1: Agent Roles and their Descriptions

LLM Integration

The system was initially integrated with state-of-the-art cloud-based multimodal LLMs, using each provider's offi-

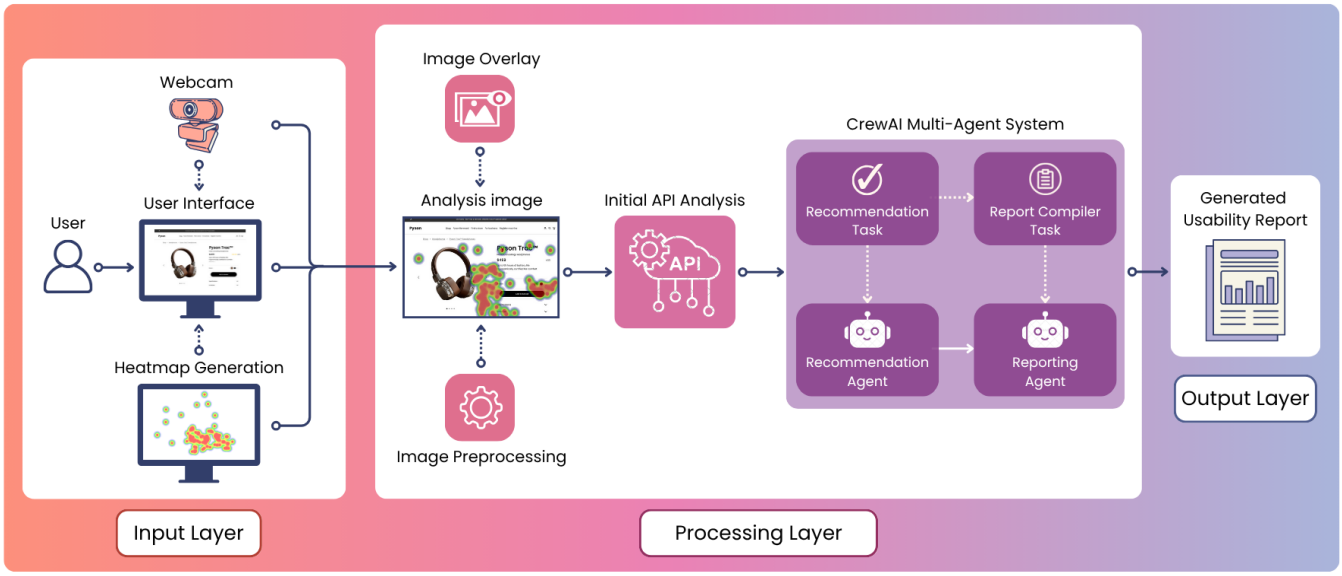

Figure 1: Overview of the Proposed Automated Usability Evaluation System

cial API to ensure reliability and compatibility. The models tested during early development can be found in Table 2.

Models	Provider
GPT-4o (OpenAI et al. 2024)	OpenAI
Gemini-2.0-Flash	Google DeepMind
Pixtral Large 2411	Mistral AI
Claude 3.5 Sonnet	Anthropic

Table 2: Multimodal Cloud-based Models Used

While these models delivered high-quality and structured outputs, they required constant internet access, which introduces accessibility barriers in low-resource environments — especially in developing regions where internet connectivity is limited, costly, or subject to data caps. To address this, the system was adapted for local deployment, enabling all processing to occur on-device without requiring internet connectivity. This supports the study's broader goal of accessibility and sustainability in usability evaluation. Local models were integrated using LMStudio, a lightweight desktop application that allows running open-source models locally via an OpenAI-compatible API. The open-source models tested for local performance can be found in Table 3.

Evaluation and Results

The evaluation of the system's output consisted of two key components. First, the generated usability reports were reviewed by the experimenter to assess their accuracy. Second, the actual interfaces provided to the participants were evaluated directly by the participants themselves. Additionally, the system was evaluated with multiple cloud-based and local LLMs capable of processing multimodal inputs

Model Name	Parameters	Provider
Gemma 3 (Team et al. 2025a)	4B	Google DeepMind
LLaVA v1.5 (Liu et al. 2023)	7B	Microsoft Research
Granite Vision (Team et al. 2025b)	3.2B	IBM Research

Table 3: Open-Source Local Multimodal Models Used

to ensure generalization. A key aspect of this evaluation involved assessing the precision of the models in classifying usability strengths and weaknesses. Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Where:

- TP **(true positives)** represents the correctly identified usability strengths or weaknesses.
- FP **(false positives)** represents cases where an aspect was misclassified as a strength or weakness.

Each model was evaluated based on:

- **Category-specific precision:** Analyzing precision separately for strengths and weaknesses to understand model-specific tendencies.
- **Runtime analysis:** Comparing the time taken by each model to generate usability reports.

Study Design and Experimental Setup

The usability study was conducted in a controlled lab environment on a personal machine with a GTX 1660 Ti GPU,

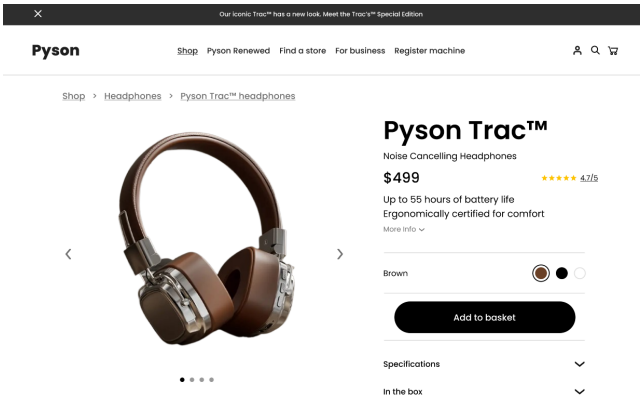

Figure 2: Good variation of the Interface

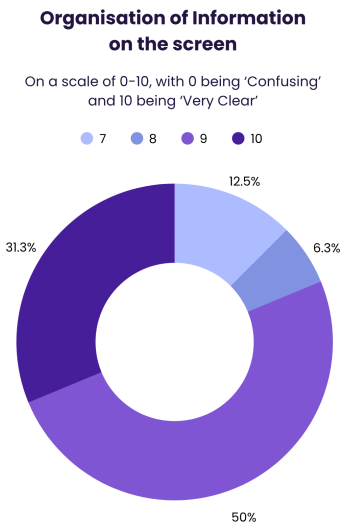

Figure 3: Participants' opinions on the Organization of Information on the screen. Question taken from (Chin, Diehl, and Norman 1988)

16GB RAM and an AMD Ryzen 7 processor. The study consisted of sixteen participants who provided their digital consent to being a part of the evaluation. They were recruited through voluntary participation without a specific statistical sampling approach. Participants were informed of the non-invasive eye-tracking setup and assured that their gaze data would be anonymized in accordance with GDPR standards. Calibration was completed using a 4×4 clickable dot grid, after which participants interacted with three UI images, each representing a variation of the same product page use case: good, moderate, and poor designs (see Figures 2, 4 and 5 respectively). Image order was randomized to reduce bias. To minimize distractions, WebGazer.js's prediction point and camera preview were hidden during the evaluation. After viewing each image, participants completed a post-image questionnaire adapted from the QUIS (Chin, Diehl, and Norman 1988) and PUTQ frameworks. This immediate feedback approach ensured higher accuracy by capturing impressions while still fresh. These UI-specific questionnaires were selected to match the study's focus on static interfaces. Questions were slightly simplified for participants with varying UI/UX expertise without altering their core intent. A final post-test questionnaire captured comparative preferences across all three UI designs.

Eye-Tracking and Usability Study Results

This subsection presents the findings from the eye-tracking and usability evaluations. First, participant responses and interaction patterns are analyzed for each interface individually. Subsequently, a comparative assessment across all three interfaces is provided to identify overarching trends and distinctions in usability performance.

Good UI Variation

While design is subjective, the good variation of the user interface, as illustrated in Figure 2, thoroughly followed industry standard and user interface guidelines such as Nielsen's Heuristics and Fitts' Law in its objective implementation. Participant data (as seen in Figure 3) supported these adherence to standards as all participants voted 7+ for the *'Organization of Information'* on a scale from 0 to 10, with 0 being *'Confusing'* and 10 being *'Very Clear'*, with half the participants voting 9 out of 10. Most participants (62.5%) did not find the interface cluttered, with only one high score attributed to human error. The system's automated analysis reinforced these findings, repeatedly identifying the visual prominence of key UI elements (e.g., product image, name, price, and CTA button) and noting contrast issues in secondary elements like breadcrumb links. The model achieved a strength precision of 97.56% and a weakness precision of 70.00%, demonstrating high reliability in identifying key usability attributes.

Moderate UI Variation

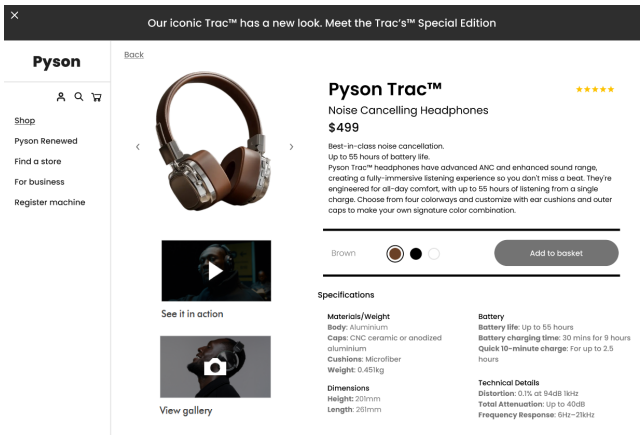

Figure 4: Moderate variation of the Interface

The moderate variation of the use case (Figure 4) was designed to introduce minor usability friction by subtly vio-

lating established UI principles while remaining functional. Unlike the good variation, which followed best practices, this version tested user adaptability to less intuitive interactions. A key issue was information clarity, only 25% of participants rated character readability at the highest level (10), while another 25% rated it at 5, indicating moderate difficulty. Scores of 8 and 9 received 12.5% each, reflecting mixed perceptions of readability. Another challenge was the visibility of interactive elements. The CTA button was intentionally blended into the background, reducing its salience. Participant feedback supported this, with only 6% selecting this variation as the easiest interface for quickly finding the primary action ('Add to basket').

The crew's analysis supports these findings, highlighting that while centrally placed product images naturally attracted attention, navigation menus and call-to-action buttons suffered from reduced engagement. Sparse fixations in key interaction zones highlighted usability issues. The model achieved 96.67% precision in identifying weaknesses but only 65.85% for strengths, indicating its stronger performance in detecting design flaws likely due to their clearer deviation from standards.

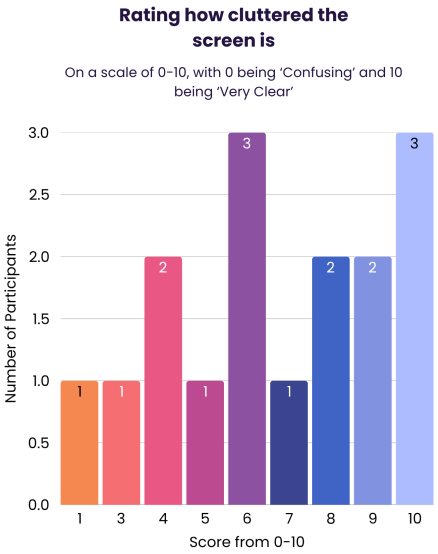

Figure 6: Participant ratings on Screen Clutter for the Bad UI. Question taken from (Chin, Diehl, and Norman 1988).

Bad UI Variation

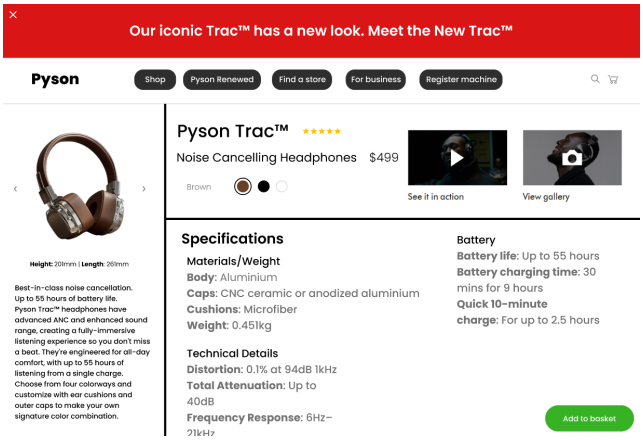

Figure 5: Bad variation of the Interface

The Bad UI variation (Figure 5) was designed to introduce severe usability issues by violating core design principles. In contrast to the good (Figure 2) and moderate (Figure 4) variations, this version intentionally disrupted layout structure and interaction flow to assess how users respond to poorly designed interfaces. A primary concern was the disorganized layout. Participant feedback reflected this clearly, reporting a Net Promoter Score (NPS) of -32 for information organization. As shown in Figure 6, several users rated the screen as highly cluttered, with three participants assigning the maximum score of 10. These results emphasize the critical role of visual structure in user navigation and satisfaction. The system's analysis supported participant feedback, revealing that while product images drew attention, key interactive elements — such as navigation menus and CTAs — suffered from low engagement due to poor contrast and layout. Sparse fixations in expected interaction zones suggested that users found the interface confusing and unpredictable. The results highlighted issues like low-contrast text, inconsistent spacing, and misaligned components, particularly noting reduced interaction with the 'See it in action' section and navigation bar. These findings stress the importance of WCAG-compliant (W3C 2023) readability and effective element positioning. The model demonstrated strong performance in detecting flaws, achieving 96.97% precision, compared to 75.61% for strengths — mirroring trends observed in the moderate UI analysis.

Comparative Analysis between Variations

The comparative evaluation of the three UI variations — Good, Moderate, and Bad — revealed significant differences in usability, as reflected in participant feedback, eye-tracking data, and system-generated analyses.

User Preferences and Feedback Trends User preferences overwhelmingly favored the Good UI, with 75% selecting it as the most usable and 94% agreeing it made the primary action ('Add to Basket') easiest to find. It also received the highest Net Promoter Score (NPS) for information organization (81), compared to the Moderate (44) and Bad UI (-32). The Bad UI was rated least clear by 56.2% of users. While the Moderate UI received mixed feedback, some appreciated the upfront product details, though satisfaction scores varied across navigation and clarity. Notably, despite poor usability performance, one user still preferred the Bad UI, and it received fewer complaints about screen clutter than the Moderate UI, suggesting some users valued information density over layout clarity.

Eye-Tracking Data and Heatmap Insights Fixation heatmaps reinforced these findings. In the Good UI, users

exhibited focused gaze clustering around essential elements such as the product image, title, rating, and 'Add to Basket' CTA. Minimal erratic scanning indicated a well-structured layout, allowing users to process information efficiently.
In the Moderate UI, increased gaze dispersion was observed, particularly around low-contrast elements, suggesting higher cognitive effort was required to locate key information.
The Bad UI demonstrated highly scattered gaze patterns, indicating a lack of visual hierarchy and excessive cognitive strain due to poor layout structuring. Users frequently shifted their focus across the page, struggling to find relevant information.

System-Generated Analysis Performance CrewAI effectively identified key usability strengths across all UI variations. In the Good UI, it achieved high precision in detecting strengths (94.12%) and moderate accuracy in spotting weaknesses (66.67%), though it sometimes misclassified contrast and secondary content issues. For the Moderate UI, it flagged contrast and spacing concerns but struggled with subtle usability inefficiencies. The Bad UI prompted accurate detection of major structural and navigation flaws. Overall, results highlight the importance of clarity and structured content, with the Good UI offering the best balance, and the Moderate UI providing insights for improving information delivery without compromising usability.

Model and Report Evaluation This evaluation was conducted by manually reviewing the output usability report and verifying the identified strengths and weaknesses against ground truth assessments of the interfaces. Additionally, system runtimes were recorded for further analysis. The evaluation was carried out across all models listed in Tables 2 and 3. During the usability study, GPT-4o was used for both API analysis and crew evaluation. For the remaining models, heatmaps from the original experiment were reused as input for both stages. This approach leveraged existing data to avoid the computational and time costs of rerunning experiments. To ensure consistency, all evaluation conditions remained constant, with only the crew model and API analysis varying across models.

Proprietary Model Results During the evaluation, notable differences were observed across the proprietary models in terms of structured output adherence and consistency in usability analysis. Claude 3.5 Sonnet struggled the most with maintaining the required structured format, often generating an analysis for only one image instead of all three. As a result, multiple experiments had to be rerun to ensure a complete evaluation. In contrast, Pixtral Large 2411 displayed a pattern of repetition, consistently identifying the same two strengths — placement of the product image and the positioning of the product name and title — across all UI variations, regardless of differences in design. This resulted in a 100% precision score in strength detection, indicating high consistency but raising concerns about adaptability to nuanced interface changes.
For weakness detection, Gemini 2.0 Flash outperformed others with 96.8% precision, followed by GPT-4o (88.2%),

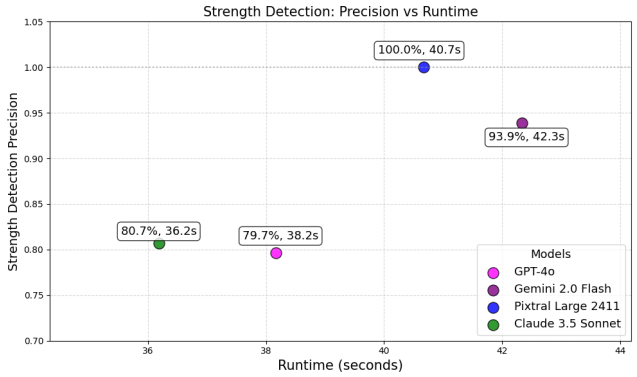

Figure 7: Precision vs Runtime Comparison of Proprietary Models for Strength Detection

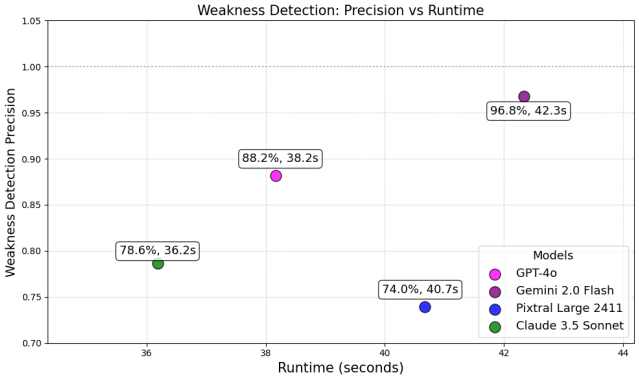

Figure 8: Precision vs Runtime Comparison of Proprietary Models for Weakness Detection

Claude 3.5 Sonnet (78.6%), and Pixtral (74.0%). While GPT-4o showed balanced performance, its slightly lower strength precision (79.7%) suggests occasional misses in identifying positive elements. Pixtral, though strong in detecting strengths, lacked variation, suggesting over-reliance on template responses. Claude, while fast, exhibited inconsistency in both structure and content. Runtime analysis showed that Gemini 2.0 Flash (42.35s) delivered the best performance overall but was the slowest. Pixtral (40.75s) achieved high strength precision through repeated patterns rather than deeper analysis. GPT-4o and Claude were the fastest (38.25s and 36.25s) but offered slightly lower precision in strengths (Figure 7). Similar trends held for weakness detection, with Gemini 2.0 Flash achieving the highest precision (Figure 8). These findings suggest a trade-off between runtime and precision. While Gemini 2.0 Flash offers superior reliability, GPT-4o balances speed and accuracy well, making it suitable for real-time evaluation. In contrast, Pixtral and Claude were either overly repetitive or structurally inconsistent, limiting their applicability in structured usability tasks. This evaluation suggests that models with longer runtimes tend to provide more accurate results, but efficiency and reliability should be considered when selecting a model for real-time usability evaluation.

Local Model Results Three open-source models were evaluated for their potential to support offline usability analysis: Gemma 3-4B, LLaVA v1.5-7B, and Granite Vision 3.2-2B. These models were tested locally via LMStudio under the same experimental conditions as proprietary models. While LLaVA v1.5-7B appeared to perform better with 52.9% strength and 54.8% weakness precision, these figures were based on just 5 successful runs out of 16 total experiments (Table 4). Many runs failed or produced unclear outputs, and in some cases, LLaVA misinterpreted the heatmap as a visual effect, exposing poor contextual understanding. Gemma 3-4B frequently returned incomplete results and achieved low precision overall (22.7%), particularly for Good UIs (3.1% weakness precision) with the longest average runtime (290.5s). Granite Vision 3.2-2B failed entirely, generating the same generic response in every case, with the lowest runtime of the three (46s). These failures likely stem from limited multimodal reasoning abilities and difficulty in handling complex visual analysis tasks. Additionally, all three models struggled with structured outputs, highlighting the current limitations of smaller open-source models for complex multimodal usability tasks.

Local Model Name	No Output	Unclear Output
Gemma 3	0	0
LLaVA v1.5	5	11
Granite Vision	0	10

Table 4: Output Completeness Across 16 Local Model Experiments

Comparative Summary Compared to proprietary models, the local open-source alternatives demonstrated significantly reduced precision, consistency, and robustness. While LLaVA showed promise in specific tasks, it lacked reliability across all UI variations. Gemma produced inconsistent outputs, and Granite failed to return usable results entirely. In contrast, proprietary models — especially Gemini 2.0 Flash and GPT-4o — provided highly structured, accurate evaluations with clear strengths in both precision and runtime balance. These results underscore the trade-off between accessibility and performance. While open-source models hold potential for offline deployment in low-resource settings, current limitations highlight the need for architectural adjustments or fine-tuning to match the reliability of proprietary LLMs. Future work should focus on improving local model integration to bridge this gap and enable sustainable, internet-independent usability evaluation workflows.

Conclusion

This study was motivated by the need to streamline UI evaluation for UI/UX practitioners by reducing the time-intensive nature of traditional usability testing. Manual assessments often require multiple iterations, limiting designers' capacity for creative problem-solving. To address this, a full-stack system was developed that integrates real-time eye-tracking with a multimodal agentic pipeline, automating the analysis process to enhance both efficiency and accuracy.

A key contribution was the transformation of qualitative feedback into quantifiable precision metrics—allowing for objective evaluation of interface effectiveness. The system outperformed prior approaches like UIClip (Wu et al. 2024), which suffered from high recall but low precision. By focusing on precision, the study reduced irrelevant suggestions and delivered more actionable feedback. Evaluation across state-of-the-art LLMs demonstrated the system's capability to accurately detect usability strengths and weaknesses. However, local deployment with smaller open-source models revealed challenges — outputs were often incomplete, lacked structure, and exhibited misinterpretations of heatmap data. This highlights the current limitations of open-source models in complex multimodal tasks.

Several technical challenges also emerged. CrewAI's limited context window made it unsuitable for direct image analysis, requiring preprocessing via an external API. Another challenge was the performance of WebGazer's prediction point (the red dot indicating gaze position), which exhibited noticeable lag. This may have been caused by the computational load of real-time heatmap generation alongside CrewAI processing. Furthermore, despite being given a structured output format, responses from CrewAI agents were not always consistent, particularly in the placement of images within the generated reports and the number of strengths/weaknesses outputted.

Future work will focus on refining the system architecture to better support smaller, locally hosted models—critical for deployment in low-resource environments where cloud-based APIs are impractical. Additionally, further optimization of the MAS will be explored to ensure its full capabilities are leveraged. There is also potential for this system to be production-ready with further refinement. The existing framework could be optimized and streamlined, improving efficiency and scalability. As the system grows in complexity, transitioning to more robust agentic frameworks like AutoGen could provide better handling of multi-agent interactions, larger context windows, and improved task orchestration.

Another key direction for improvement is extending the system to work with interactive prototypes rather than just static UI images. While static UI evaluation provides valuable insights, integrating usability testing with functional interfaces would significantly enhance the system's applicability and accuracy, making it even more practical for real-world UI/UX evaluation.

References

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *ArXiv*, abs/2302.04023.

Bertão, R. A.; and Joo, J. 2021. Artificial intelligence in

UX/UI design: a survey on current adoption and [future] practices. *Blucher Design Proceedings*.

Buschek, D.; Anlauff, C.; and Lachner, F. 2020. Paper2Wire – A Case Study of User-Centred Development of Machine Learning Tools for UX Designers. *i-com*, 20: 19 – 32.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Qian, C.; Chan, C.-M.; Qin, Y.; Lu, Y.-T.; Xie, R.; Liu, Z.; Sun, M.; and Zhou, J. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors in Agents. *ArXiv*, abs/2308.10848.

Chin, J. P.; Diehl, V. A.; and Norman, K. L. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In *International Conference on Human Factors in Computing Systems*.

Dohan, D.; Xu, W.; Lewkowycz, A.; Austin, J.; Bieber, D.; Lopes, R. G.; Wu, Y.; Michalewski, H.; Saurous, R. A.; Sohl-Dickstein, J. N.; Murphy, K.; and Sutton, C. 2022. Language Model Cascades. *ArXiv*, abs/2207.10342.

Duan, P.; Warner, J.; Li, Y.; and Hartmann, B. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Gajos, K. Z.; and Weld, D. S. 2005. Preference elicitation for interface optimization. *Proceedings of the 18th annual ACM symposium on User interface software and technology*.

Gardey, J. C.; Grigera, J.; Rodríguez, A.; Rossi, G.; and Garrido, A. 2022. Predicting interaction effort in web interface widgets. *Int. J. Hum. Comput. Stud.*, 168: 102919.

Ge, Y.; Zhao, S.; Zeng, Z.; Ge, Y.; Li, C.; Wang, X.; and Shan, Y. 2023. Making LLaMA SEE and Draw with SEED Tokenizer. *ArXiv*, abs/2310.01218.

Haddad, S.; Latifzadeh, K.; Duraisamy, S.; Vanderdonckt, J.; Dâassi, O.; Belghith, S.; and Leiva, L. A. 2024. Good GUIs, Bad GUIs: Affective Evaluation of Graphical User Interfaces. *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*.

Hasse, C.; and Bruder, C. 2015. Eye-tracking measurements and their link to a normative model of monitoring behaviour. *Ergonomics*, 58: 355 – 367.

Hong, S.; Zheng, X.; Chen, J. P.; Cheng, Y.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z. H.; Zhou, L.; Ran, C.; Xiao, L.; and Wu, C. 2023. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. *ArXiv*, abs/2308.00352.

Huang, R.; Huang, J.-B.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. In *International Conference on Machine Learning*.

Jacob, R. J. K. 2002. Eye tracking in human-computer interaction and usability research : Ready to deliver the promises.

Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society. *ArXiv*, abs/2303.17760.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved Baselines with Visual Instruction Tuning.

Miniukovich, A.; and Angeli, A. D. 2015. Computation of Interface Aesthetics. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Nielsen, J. 1994. Enhancing the explanatory power of usability heuristics. *Conference Companion on Human Factors in Computing Systems*.

Novák, J. .; Masner, J.; Benda, P.; Simek, P.; and Merunka, V. 2023. Eye Tracking, Usability, and User Experience: A Systematic Review. *Int. J. Hum. Comput. Interact.*, 40: 4484–4500.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker,

S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pandian, V. P. S.; Suleri, S.; Beecks, C.; and Jarke, M. 2020. MetaMorph: AI Assistance to Transform Lo-Fi Sketches to Higher Fidelities. *Proceedings of the 32nd Australian Conference on Human-Computer Interaction*.

Papoutsaki, A.; Sangkloy, P.; Laskey, J.; Daskalova, N.; Huang, J.; and Hays, J. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *International Joint Conference on Artificial Intelligence*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Poole, A.; and Ball, L. J. 2004. Eye Tracking in Human-Computer Interaction and Usability Research : Current Status and Future Prospects.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Shen, L.; Li, H.; Wang, Y.; and Qu, H. 2024. From Data to Story: Towards Automatic Animated Data Video Creation with LLM-based Multi-Agent Systems. *ArXiv*, abs/2408.03876.

Shneiderman, B. 1998. Designing the User Interface: Strategies for Effective Human-Computer Interaction.

Stige, Å.; Zamani, E. D.; Mikalef, P.; and Zhu, Y. 2023. Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda. *Inf. Technol. People*, 37: 2324–2352.

Sun, Q.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; Wang, Y.; Gao, H.; Liu, J.; Huang, T.; and Wang, X. 2023. Generative Pretraining in Multimodality. *ArXiv*, abs/2307.05222.

Talebirad, Y.; and Nadiri, A. 2023. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. *ArXiv*, abs/2306.03314.

Tang, Z.; Yang, Z.; Khademi, M.; Liu, Y.; Zhu, C.; and Bansal, M. 2023a. CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27415–27424.

Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; and Bansal, M. 2023b. Any-to-Any Generation via Composable Diffusion. *ArXiv*, abs/2305.11846.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; Rouillard, L.; Mesnard, T.; Cideron, G.; bastien Grill, J.; Ramos, S.; Yvinec, E.; Casbon, M.; Pot, E.; Penchev, I.; Liu, G.; Visin, F.; Kenealy, K.; Beyer, L.; Zhai, X.; Tsitsulin, A.; Busa-Fekete, R.; Feng, A.; Sachdeva, N.; Coleman, B.; Gao, Y.; Mustafa, B.; Barr, I.; Parisotto, E.; Tian, D.; Eyal, M.; Cherry, C.; Peter, J.-T.; Sinopalnikov, D.; Bhupatiraju, S.; Agarwal, R.; Kazemi, M.; Malkin, D.; Kumar, R.; Vilar, D.; Brusilovsky, I.; Luo, J.; Steiner, A.; Friesen, A.; Sharma, A.; Sharma, A.; Gilady, A. M.; Goedeckemeyer, A.; Saade, A.; Feng, A.; Kolesnikov, A.; Bendebury, A.; Abdagic, A.; Vadi, A.; György, A.; Pinto, A. S.; Das, A.; Bapna, A.; Miech, A.; Yang, A.; Paterson, A.; Shenoy, A.; Chakrabarti, A.; Piot, B.; Wu, B.; Shahriari, B.; Petrini, B.; Chen, C.; Lan, C. L.; Choquette-Choo, C. A.; Carey, C.; Brick, C.; Deutsch, D.; Eisenbud, D.; Cattle, D.; Cheng, D.; Paparas, D.; Sreepathihalli, D. S.; Reid, D.; Tran, D.; Zelle, D.; Noland, E.; Huizenga, E.; Kharitonov, E.; Liu, F.; Amirkhanyan, G.; Cameron, G.; Hashemi, H.; Klimczak-Plucińska, H.; Singh, H.; Mehta, H.; Lehri, H. T.; Hazimeh, H.; Ballantyne, I.; Szpektor, I.; Nardini, I.; Pouget-Abadie, J.; Chan, J.; Stanton, J.; Wieting, J.; Lai, J.; Orbay, J.; Fernandez, J.; Newlan, J.; yeong Ji, J.; Singh, J.; Black, K.; Yu, K.; Hui, K.; Vodrahalli, K.; Greff, K.; Qiu, L.; Valentine, M.; Coelho, M.; Ritter, M.; Hoffman, M.; Watson, M.; Chaturvedi, M.; Moynihan, M.; Ma, M.; Babar, N.; Noy, N.; Byrd, N.; Roy, N.; Momchev, N.; Chauhan, N.; Sachdeva, N.; Bunyan, O.; Botarda, P.; Caron, P.; Rubenstein, P. K.; Culliton, P.; Schmid, P.; Sessa, P. G.; Xu, P.; Stanczyk, P.; Tafti, P.; Shivanna, R.; Wu, R.; Pan, R.; Rokni, R.; Willoughby, R.; Vallu, R.; Mullins, R.; Jerome, S.; Smoot, S.; Girgin, S.; Iqbal, S.; Reddy, S.; Sheth, S.; Põder, S.; Bhatnagar, S.; Panyam, S. R.; Eiger, S.; Zhang, S.; Liu, T.; Yacovone, T.; Liechty, T.; Kalra, U.; Evci, U.; Misra, V.; Roseberry, V.; Feinberg, V.; Kolesnikov, V.; Han, W.; Kwon, W.; Chen, X.; Chow, Y.; Zhu, Y.; Wei, Z.; Egyed, Z.; Cotruta, V.; Giang, M.; Kirk, P.; Rao, A.; Black, K.; Babar, N.; Lo, J.; Moreira, E.; Martins, L. G.; Sanseviero, O.; Gonzalez, L.; Gleicher, Z.; Warkentin, T.; Mirrokni, V.; Senter, E.; Collins, E.; Barral, J.; Ghahramani, Z.; Hadsell, R.; Matias, Y.; Sculley, D.; Petrov, S.; Fiedel, N.; Shazeer, N.; Vinyals, O.; Dean, J.; Hassabis, D.; Kavukcuoglu, K.; Farabet, C.; Buchatskaya, E.; Alayrac, J.-B.; Anil, R.; Dmitry; Lepikhin; Borgeaud, S.; Bachem, O.; Joulin, A.; Andreev, A.; Hardin, C.; Dadashi, R.; and Hussenot, L. 2025a. Gemma 3 Technical Report. arXiv:2503.19786.

Team, G. V.; Karlinsky, L.; Arbelle, A.; Daniels, A.; Nassar, A.; Alfassi, A.; Wu, B.; Schwartz, E.; Joshi, D.; Kondic, J.; Shabtay, N.; Li, P.; Herzig, R.; Abedin, S.; Perek, S.; Harary, S.; Barzelay, U.; Goldfarb, A. R.; Oliva, A.; Wieles, B.; Bhattacharjee, B.; Huang, B.; Auer, C.; Gutfreund, D.; Beymer, D.; Wood, D.; Kuehne, H.; Hansen, J.; Shtok, J.; Wong, K.; Bathen, L. A.; Mishra, M.; Lysak, M.; Dolfi, M.; Yurochkin, M.; Livathinos, N.; Harel, N.; Azulai, O.; Na-

parstek, O.; de Lima, R. T.; Panda, R.; Doveh, S.; Gupta, S.; Das, S.; Zawad, S.; Kim, Y.; He, Z.; Brooks, A.; Goodhart, G.; Govindjee, A.; Leist, D.; Ibrahim, I.; Soffer, A.; Cox, D.; Soule, K.; Lastras, L.; Desai, N.; Ofek-koifman, S.; Raghavan, S.; Syeda-Mahmood, T.; Staar, P.; Drory, T.; and Feris, R. 2025b. Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence. arXiv:2502.09927.

van den Berg, L.; Engelsma, T.; and Peute, L. 2024. Exploration of Eye-Tracking Methodologies in Usability Testing of Digital Health Technology: A Rapid Review. *Studies in health technology and informatics*, 316: 1130–1134.

W3C. 2023. Web Content Accessibility Guidelines (WCAG) 2.2. Accessed: 17-Oct-2024.

Wang, J.; Antonenko, P. D.; Celepkolu, M.; Jimenez, Y.; Fieldman, E.; and Fieldman, A. 2019. Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data. *International Journal of Human–Computer Interaction*, 35: 483 – 494.

Wu, J.; Peng, Y.-H.; Li, X. Y. A.; Swearngin, A.; Bigham, J. P.; and Nichols, J. 2024. UIClip: A Data-driven Model for Assessing User Interface Design. In *ACM Symposium on User Interface Software and Technology*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Zhang, S.; Zhu, E.; Li, B.; Jiang, L.; Zhang, X.; and Wang, C. 2023a. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *ArXiv*, abs/2308.08155.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2023b. NExT-GPT: Any-to-Any Multimodal LLM. *ArXiv*, abs/2309.05519.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *ArXiv*, abs/1504.06755.

Zelinskyi, S.; and Boyko, Y. 2024. Integrating session recording and eye-tracking: development and evaluation of a Chrome extension for user behavior analysis. *Radioelectronic and Computer Systems*.

Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; Yan, H.; Fu, J.; Gui, T.; Sun, T.; Jiang, Y.; and Qiu, X. 2024. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. *ArXiv*, abs/2402.12226.