Seeing Twice: How Side-by-Side T2I Comparison Changes Auditing Strategies

MATHEUS KUNZLER MALDANER, University of Florida, USA
WESLEY HANWEN DENG, Carnegie Mellon University, USA
JASON I. HONG, Carnegie Mellon University, USA
KEN HOLSTEIN, Carnegie Mellon University, USA
MOTAHHARE ESLAMI, Carnegie Mellon University, USA

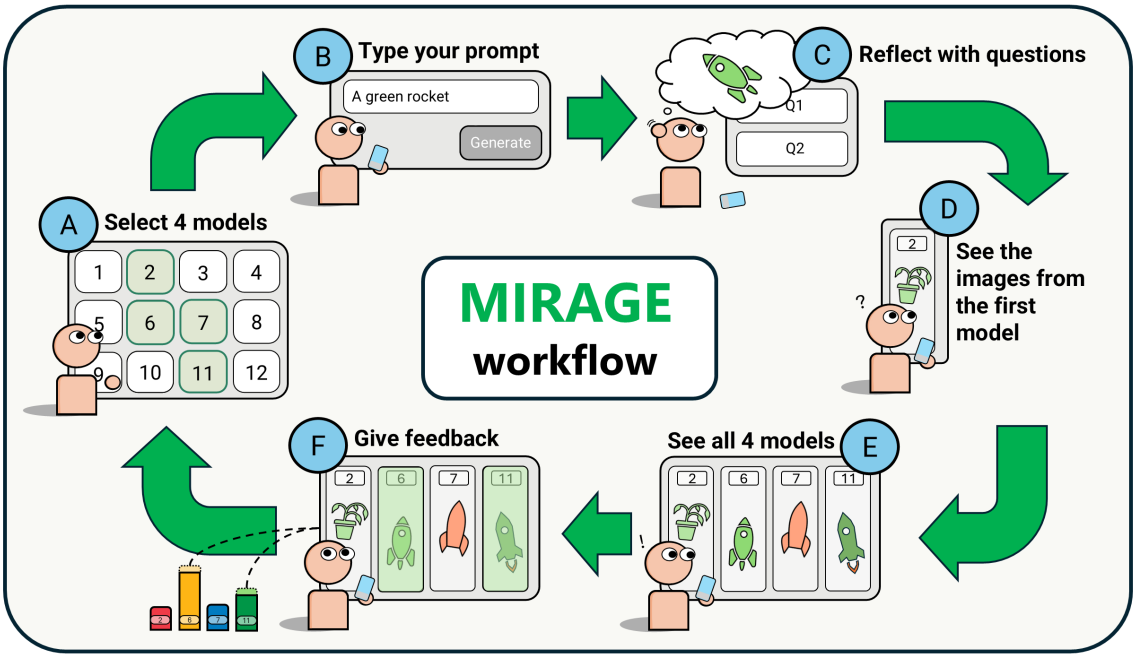

Fig. 1. MIRAGE supports the comparison of different text-to-image models displayed side-by-side in a single window. Users can select up to four models to be displayed (A) and enter a prompt of their choice (B). While the generation happens in the background, users are asked to reflect on what they expect to see (C, Q1) and why they chose that specific prompt (C, Q2). Users in this study first saw the outputs of just one of the four models (D) and then of all of them (E). Finally, the users can analyze all of the images and select which models matched their original expectations while pointing out any details they might have not noticed before (F).

While generative AI systems have gained popularity in diverse applications, their potential to produce harmful outputs limits their trustworthiness and utility. A small but growing line of research has explored tools and processes to better engage non-AI expert users in auditing generative AI systems. In this work, we present the design and evaluation of MIRAGE, a web-based tool exploring a "contrast-first" workflow that allows users to pick up to four different text-to-image (T2I) models, view their images side-by-side, and provide feedback on model performance on a single screen. In our user study with fifteen participants, we used four predefined models for consistency, with only a single model initially being shown. We found that most participants shifted from analyzing individual images to general model output patterns once the side-by-side step appeared with all four models; several participants

Authors' Contact Information: Matheus Kunzler Maldaner, mkunzlermaldaner@ufl.edu, University of Florida, Gainesville, Florida, USA; Wesley Hanwen Deng, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Jason I. Hong, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Ken Holstein, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Motahhare Eslami, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

coined persistent "model personalities" (e.g., *cartoonish*, *saturated*) that helped them form expectations about how each model would behave on future prompts. Bilingual participants also surfaced a language-fidelity gap, as English prompts produced more accurate images than Portuguese or Chinese, an issue often overlooked when dealing with a single model. These findings suggest that simple comparative interfaces can accelerate bias discovery and reshape how people think about generative models.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning approaches**; • **Social and professional topics** → *Cultural characteristics*; • **Software and its engineering** → *Software creation and management*.

Additional Key Words and Phrases: Generative AI, Text-to-image Models, AI Auditing, Crowdsourcing, Interactive System Design

1 Introduction

Auditing is crucial to ensure fairness in AI systems [14]. At a high level, AI auditing refers to a process of repeatedly testing an algorithm with inputs and observing the corresponding outputs, in order to understand its behavior and potential external impacts [2, 11]. Recognizing the power of diverse end users in surfacing harmful behaviors in AI systems that might otherwise be overlooked by small, homogeneous groups of AI developers, recent research has explored the potential of engaging users in auditing AI systems [5, 7, 15]. Some platforms, such as Hugging Face and OpenAI, have developed interfaces using side-by-side comparisons for GenAI outputs, as a method to help users or red-teamers review and audit GenAI outputs [3]. However, there remains a lack of scientific understanding of how side-by-side comparison can support activities such as AI auditing and red-teaming, as well as other potential human-AI interactions.

In this paper, we explore whether **viewing outputs from multiple text-to-image models side-by-side can help users identify details they might miss when only looking at a single model's output.** We extend our previous work on (1) the design and development of MIRAGE[1] [10], a web application supporting multi-model side-by-side comparison while soliciting users' perception changes, (2) an expansion of the findings from a study with fifteen users, and (3) the exploration of future work to leverage the MIRAGE framework to further explore the effects of side-by-side comparison for AI auditing and beyond.

2 Related Work

A small but growing line of work has explored developing tools for engaging AI developers and users in testing and auditing AI models. Current examples include side-by-side comparators such as IndieLabel, LLM-Comparator, and WeAudit [6, 8, 9], and industry dashboards from Hugging Face, OpenAI, Replicate, and Chatbot Arena for red-teaming [3, 4, 13, 16, 17]. This paper extends prior work by developing a web-based interface to explore how reviewing outputs from multiple *distinct* T2I models simultaneously affects users' abilities and strategies of conducting AI audits.

3 System Overview

MIRAGE provides a simple workflow where users select up to four models from the thirty available models, enter a prompt, and see the outputs side-by-side. Each model generates eight images, for a total of thirty-two images across all models. For this study, the same four models were fixed across sessions. The new MIRAGE extension codebase has been made modular, enabling new models to be added with minimal effort.

To mask latency, MIRAGE employs "micro-interaction" buffering [1]. While participants answer two reflection questions, images render in the background, so most results appear instantly when participants look back [12]. Several

[1] MIRAGE stands for **M**ulti-model **I**nterface for **R**eviewing and **A**uditing **Ge**nerative Text-to-Image AI

participants even assumed the images were pre-cached rather than generated live. The implementation, deployment, and interface details appear in Appendix A.2 and Appendix A.3. Please visit mirage.weaudit.org to access MIRAGE [2].

4 Pilot User Study

Fifteen volunteers completed a 45-minute, IRB-approved session and were compensated with a $15 gift card. Each session started with a predefined prompt ("a fancy dinner party"), and users then contributed two self-chosen prompts based on their lived experiences. While MIRAGE worked in the background, users answered two questions ("Why did you choose this prompt?" and "What do you expect to see?"), giving the system time to generate unseen outputs from four fixed T2I models: SDXL-Lightning, Kandinsky 2.2, Stable Diffusion, and Latent Consistency. To reduce latency bias, SDXL-Lightning's outputs were always displayed first because of its generation speed. After rating that single-model view, participants saw the full four-panel grid (Fig. 2), noted any newly observed details, and described how their auditing strategy shifted. Six sessions were conducted in Portuguese or Chinese, providing an early look at multilingual prompts. Full demographic details appear in Appendix A.1.

5 Preliminary Findings

Overall, participants reported that MIRAGE was generally clear and intuitive to use. Some also appreciated the question flow (see Figure 2), which helped them reflect on their own biases and articulate the reasoning behind their prompt choices during auditing. In the sections below, we expand on two key insights into how MIRAGE inspired users to develop new auditing strategies both in terms of low-level model output details and high-level model "personalities," expanding the affordances offered by prior user-AI auditing tools [6, 9].

5.1 Auditing Strategies Inspired by Contrasting Multi Model Outputs

The exposure to multiple images led participants to develop new auditing strategies. In particular, P1, P4, P5, and P13 reported that they focused on inspecting individual images when there was only one model but focused on reviewing the overall image output distribution when there were multiple models. When asked how seeing more model outputs affected their strategies for analyzing the images, participants noted a clear shift in their approach. P4 specifically mentioned that with multiple models, they were *"less likely to look at individual images and more at surface-level outputs*

[2]We plan to live demo MIRAGE during the conference

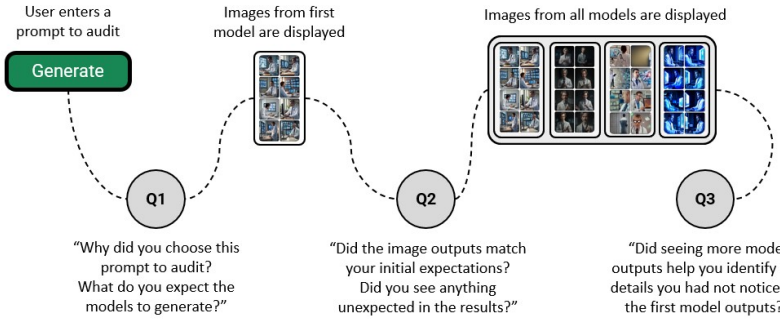

Fig. 2. MIRAGE User Study Workflow

and trends," emphasizing a more holistic approach. P11 and P14 described how multi-model viewing helped them notice details they had missed in single-model outputs, such as biases in age, color schemes, or stylistic consistencies, with P11 remarking that they "had not noticed how good the first model was until they saw the others."

In addition, viewing outputs from additional models enabled participants to identify new image details that could lead to potentially harmful biases that were not recognized when viewing the outputs of a single model. For instance, when P1 saw the first model's output for the prompt "An intern working in the city of Pittsburgh," they only noted that all workers were depicted outdoors. However, after seeing additional model outputs, they observed that darker-skinned workers were depicted wearing construction clothes, while lighter-skinned workers were depicted in office attire, carrying keyboards and notebooks. In another example, participants who initially expected the prompt "A fancy dinner party" to generate images of white males noticed that all tables were long and rectangular, a more Western style, as opposed to the round tables commonly found in Asian cultures.

5.2 Exploring Model Personalities

In addition to assisting participants in auditing AI-generated images, we also found that side-by-side comparison helped participants compare and understand model output styles, or the model "personalities." For example, several participants (P5, P7, P8, P11, P10) praised ByteDance Sdxl as offering the "best quality" and "most variation," though P4 found its outputs "least culturally diverse." Latent Consistency elicited comments like "cartoonish" (P5) or "exaggerated" (P3), and some (P4, P6, P7) noted a lack of diversity in the outputs. Interestingly, P7 said that its consistency was beneficial for generating coherent storylines. Kandinsky was described as "colorful and vibrant" (P8, P5, P10). Finally, participants found Stable Diffusion's proportions "off" (P3, P5, P6, P8, P11, P10), but P7 felt it generally "matched expectations." These varied impressions underscore that each model may cater to different user needs, prompting future inquiry into how style, realism, or diversity impacts user preference.

6 Future Work

Looking ahead, we propose the following future use cases with the goal of bridging the gap between everyday users and developers.

(1) Controlled Experiments: Run large-scale controlled experiments to *quantify* how side-by-side workflows improve bias detection and how factors such as identity or AI literacy mediate those gains.

(2) Anonymous Auditing: Add an *anonymous-auditing* mode so developers can submit proprietary models that are anonymized in MIRAGE, letting diverse users surface harms without exposing trade secrets.

(3) Model Leaderboard: Launch a community *leaderboard* that ranks T2I systems by real-user preference clicks, creating a feedback loop that motivates developers to fix issues highlighted through MIRAGE.

Through collaboration between users from varying backgrounds and AI developers, MIRAGE has the potential to drive meaningful advancements in AI transparency and auditing.

References

[1] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 313–322.

[2] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 612–643.

[3] Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, Roman Lutz, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Eugenia Kim, Justin Song, Keegan Hines, Daniel Jones, Giorgio

Severi, Richard Lundeen, Sam Vaughan, Victoria Westerhoff, Pete Bryan, Ram Shankar Siva Kumar, Yonatan Zunger, Chang Kawaguchi, and Mark Russinovich. 2025. Lessons From Red Teaming 100 Generative AI Products. arXiv:2501.07238 [cs.AI] https://arxiv.org/abs/2501.07238

[4] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs.AI] https://arxiv.org/abs/2403.04132

[5] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[6] Wesley Hanwen Deng, Claire Wang, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *arXiv preprint arXiv:2501.01397* (2025).

[7] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[8] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2024. LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. arXiv:2402.10524 [cs.HC] https://arxiv.org/abs/2402.10524

[9] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (nov 2022), 34 pages. doi:10.1145/3555625

[10] Matheus Kunzler Maldaner, Wesley Hanwen Deng, Jason Hong, Ken Holstein, and Motahhare Eslami. 2024. MIRAGE: Multi-model Interface for Reviewing and Auditing Generative Text-to-Image AI. In *HCOMP 2024 Works-in-Progress: The Twelfth AAAI Conference on Human Computation and Crowdsourcing*. Pittsburgh, PA, USA. https://www.humancomputation.com/assets/wip_2024/HCOMP_24_WIP_4.pdf

[11] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[12] Fiona Fui-Hoon Nah. 2004. A study on tolerable waiting time: how long are Web users willing to wait? *Behaviour & Information Technology* 23, 3 (2004), 153–163. doi:10.1080/01449290410001669914

[13] Replicate. 2024. Replicate: Run AI with an API. https://replicate.com/

[14] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. https://api.semanticscholar.org/CorpusID:15686114

[15] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (oct 2021), 29 pages. doi:10.1145/3479577

[16] TestingDocs.com. 2024. Compare OpenAI Models Side-By-Side. https://www.testingdocs.com/compare-openai-models-side-by-side/ Accessed: 2025-01-23.

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

A Appendix

A.1 Participant Demographics

The participants in our pilot study encompassed a group of fifteen individuals aged 18 to 76, spanning ethnicities, educational backgrounds, and genders. Participants were recruited through convenience sampling, and the group included native speakers of English, Portuguese, and Chinese, which enabled the exploration of cross-linguistic and cultural differences in text-to-image (T2I) model auditing. Participants' experience using artificial intelligence and T2I models and tools varied widely, with proficiency levels ranging from minimal exposure (1) to advanced expertise (5) on a 1–5 scale. This diversity allowed us to evaluate the usability and effectiveness of MIRAGE across a broad spectrum of users, incorporating varied cultural contexts, technical expertise, and lived experiences, as summarized in Table 1.

A.2 Technical Implementation

In order to achieve our goal, we first developed the MIRAGE web application (Figure 3), and hosted it on Amazon Lightsail with a Docker container packaging a Django project. Image generation for specific models is handled through

Participant	Age	Gender	Ethnicity	Educational Level	Language	AI Exp.	T2I Exp.
P01	20	Female	Asian	Bachelor's degree	English	3	3
P02	21	Male	Indian	Bachelor's degree	English	5	3
P03	19	Female	Black	Bachelor's degree	English	3	1
P04	22	Female	Asian	Master's degree	English	2	1
P05	18	Female	Latino, White	High School	English	3	2
P06	22	Male	Latino, White	Bachelor's degree	Portuguese	3	2
P07	76	Male	Latino, White	Master's degree	Portuguese	1	1
P08	63	Male	Latino, White	Bachelor's degree	Portuguese	5	5
P09	25	Male	Asian	Bachelor's degree	Chinese, English	2	1
P10	20	Female	Asian	Bachelor's degree	English	3	4
P11	25	Female	Asian	Master's degree	English	3	3
P12	45	Female	Asian	High School	Chinese, English	3	1
P13	20	Male	Asian	Bachelor's degree	English	4	3
P14	50	Female	Latino, White	Bachelor's degree	Portuguese	1	1
P15	21	Female	White	Bachelor's degree	English	5	4

Table 1. Demographics and AI experience of participants in the MIRAGE pilot study.

API calls to Replicate, a centralized service that works as a hub for open-source models. Replicate stores the generated images in their servers for up to an hour, so we save all images to our own Amazon S3 Bucket with unique IDs for persistence. The images and their associated audits are referenced in an Amazon DynamoDB table for easy retrieval and analysis. Backend computations, including calls to Replicate, are performed using AWS Lambda functions, accessed through Amazon API Gateway. Although not used in our study, the Discourse API integration is designed to facilitate community discussions about audit findings in future iterations of MIRAGE. Lastly, the design of the system is modular, allowing new T2I models to be added as they become available. While users are encouraged to try the system deployed at mirage.weaudit.org, we acknowledge that research teams may want to explore different research questions using MIRAGE. We therefore plan to open-source the application in the coming months.

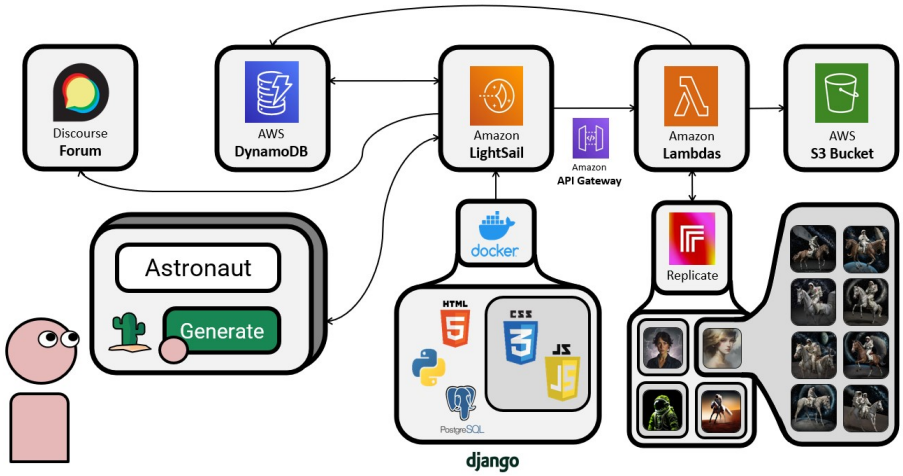

Fig. 3. MIRAGE Technical Implementation.

A.3 Application Interface

In order to reduce friction, we sought to create a minimalist and straightforward application. All users described MIRAGE as very easy to use. Figure 4 shows the overall interface of our tool, broken down into three clearly labeled main areas. As seen in Figure 6, we are open to tailoring MIRAGE for specific research studies. Figure 5 shows a "Portuguese" option for a customized version of MIRAGE designed specifically for a Brazilian research group.

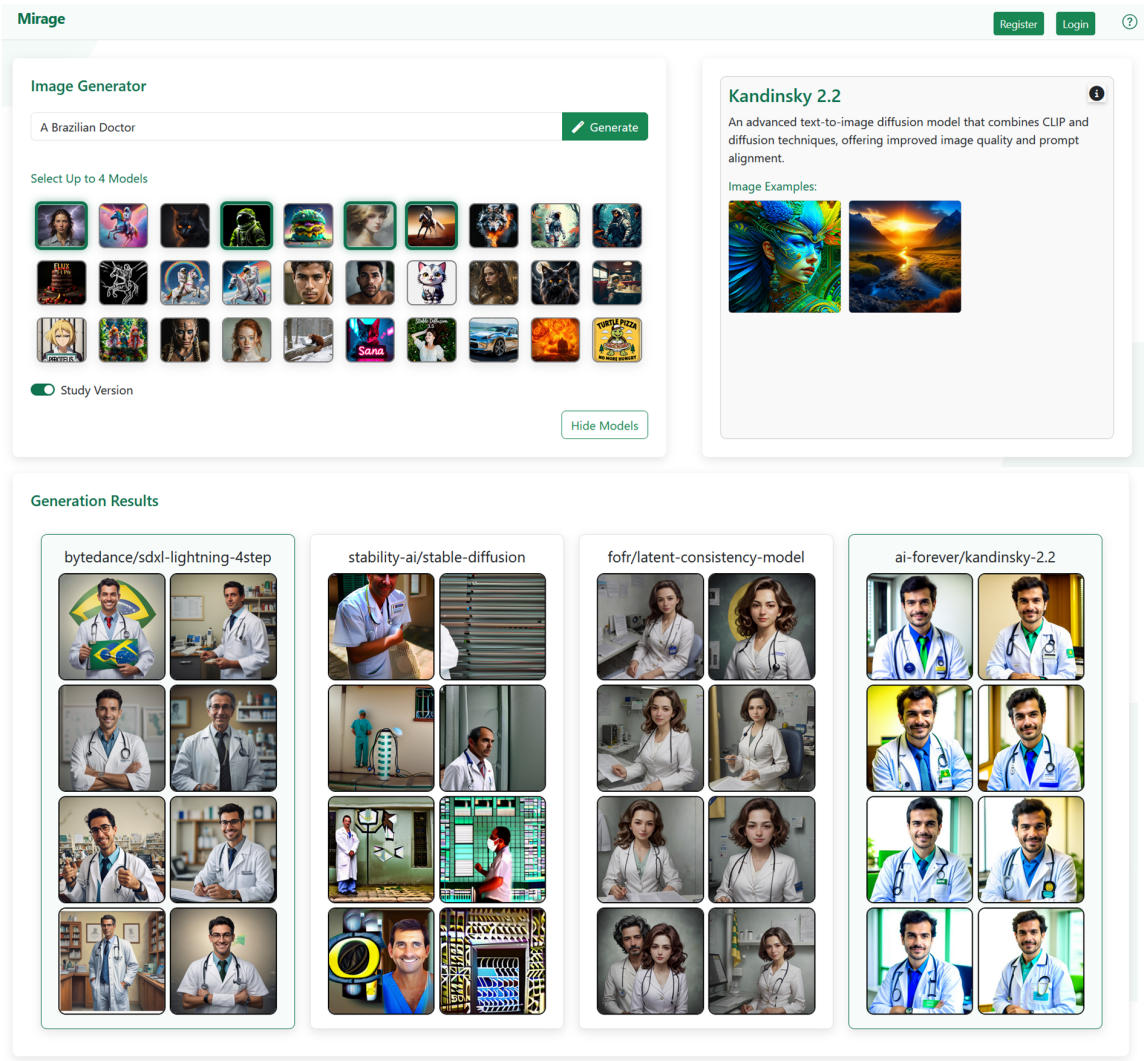

Fig. 4. MIRAGE Main Page. Image Generator (upper-left) lets users type a prompt, toggle the "Study Version" option for the user study scenario, and pick up to four models from a total of thirty thumbnail icons. Hovering over a model opens the Model Preview panel (upper-right), which shows example images. The upper-right panel is also populated with reflection questions during the study workflow. Pressing Generate fills the Generation Results area with eight images per model, arranged in four parallel columns, so users can compare thirty-two outputs at a glance.

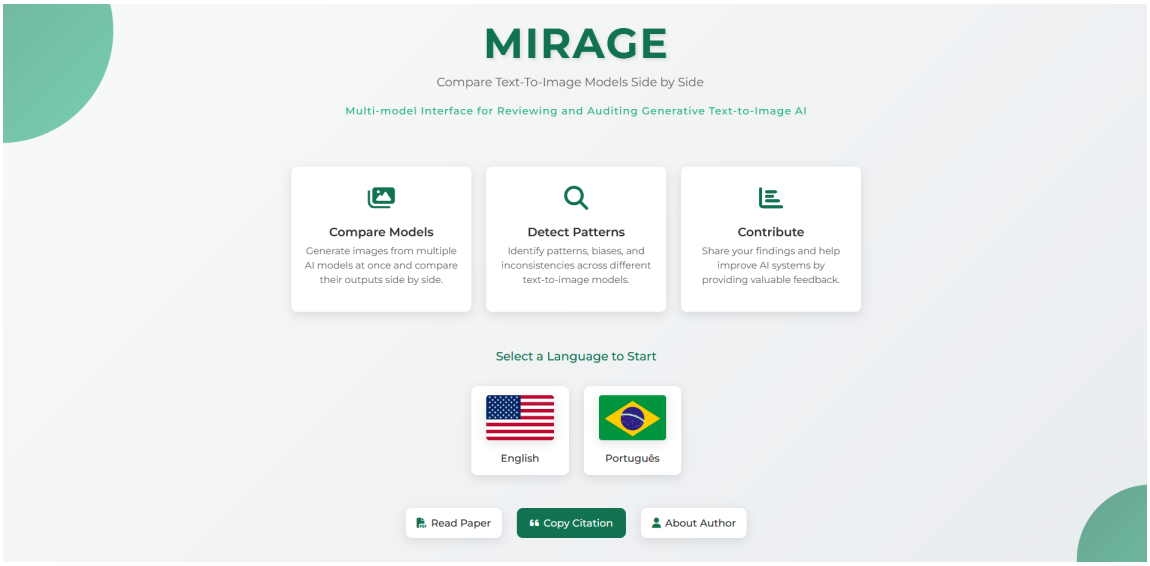

Fig. 5. MIRAGE Landing Page. Users can read the purpose of the tool, select a language and read previously published work.

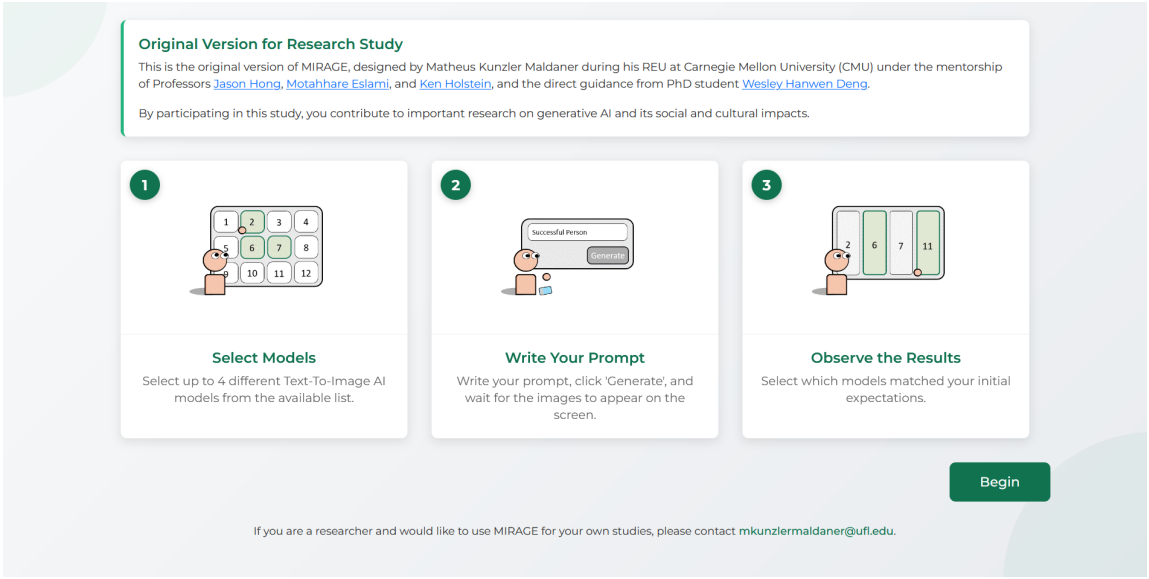

Fig. 6. MIRAGE Instructions Page. Users are shown a simple three-step approach to using the tool.