

Mirage Deepfake Detection Agent – Market Research Report

Market Size, Share, and Product Fit

The deepfake detection market is rapidly growing in response to the explosion of AI-generated fake content. In 2023 the global deepfake **detection** market was valued at only about \$213 million, but it is forecast to surge to roughly **\$3.46 billion by 2031** – a **40%+ CAGR** growth trajectory. This explosive growth is driven by the rising threat of deepfake fraud and misinformation across industries. For example, identity fraud attacks using deepfakes jumped 31× from 2022 to 2023 in one report, underscoring the urgent need for reliable detection. Major sectors are already seeking solutions: banking and financial services want to prevent deepfake identity/document fraud; media and entertainment companies need tools to **preserve trust and authenticity** by stopping the spread of fake videos and images; governments and defense agencies view deepfakes as a national security risk; and social media platforms are exploring anti-deepfake tech to maintain a safe environment for users. **Journalists and fact-checkers** in particular are looking for ways to verify images, videos, or audio clips, as public trust in media is eroded by convincing fake visuals. This is where *Mirage* fits in: it targets a clear pain-point across these domains by offering a multi-modal deepfake detection agent (covering text, images, and video) that delivers a “genuineness score” for content. Its use of **blockchain (NEAR Protocol)** for immutable logging of verification results is a unique differentiator, adding transparency and public trust to the detection (similar in spirit to initiatives like Project Origin which use tamper-proof ledgers to certify media authenticity). Likewise, leveraging a **TEE (Trusted Execution Environment)** for the AI agent addresses security and privacy concerns, ensuring sensitive analyses are done in a private, verifiable manner. Overall, *Mirage* is well-positioned to serve its target audiences – enterprises, social media platforms, and journalists – by filling the market need for a **trustworthy, real-time deepfake detection** solution that is both **highly accurate** and **credibly transparent** (thanks to on-chain verification).

Competitors in Deepfake Detection

The emerging deepfake detection industry has a mix of specialized startups and tech giants developing solutions. A recent industry report highlights key players including **Microsoft, WeVerify, Resemble AI, Deepgram, Pindrop, Amped Software, Intel, SentinelOne, DeepWare AI, Sensity AI, and Reality Defender**. Below are some notable competitors and solutions similar to *Mirage*:

- **Sensity AI** – A leading deepfake detection company specializing in monitoring online platforms for fake media. Sensity’s tools scan vast amounts of content in real time,

flagging suspicious videos or images and helping combat the spread of harmful deepfakes on social media. They offer enterprise solutions (not consumer-free) with advanced machine learning to recognize AI-generated visuals.

- **Reality Defender** – A deepfake detection platform that employs sophisticated algorithms for real-time monitoring. Reality Defender provides enterprises with actionable alerts to mitigate risks from fraudulent media, focusing on quick detection to stop deepfakes before they cause harm. They often integrate into corporate security workflows and claim high accuracy in spotting manipulated videos.
- **Deepware** – An open-source deepfake video detector that allows users to upload videos to an online scanner for analysis. Deepware's technology looks for manipulation signs in faces within videos. They also provide an API and on-premise solutions for cybersecurity use. However, its current limitation (as noted by journalists) is difficulty detecting AI **voice** swaps in videos – it's mainly effective for face-swap deepfakes.
- **Hive Moderation (Hive AI)** – A broader AI content detection service that includes deepfake detection for images **and text**. Hive offers a free web-based scanner and even a Chrome extension to let users quickly check if an image or text was AI-generated. This lowers the barrier for independent fact-checkers to verify content. Hive's solution is fast and user-friendly, exemplifying features (like browser integration) that users appreciate.
- **Illuminarty** – A newer tool that detects AI-generated images and text. Illuminarty stands out by not only flagging AI-created images but also identifying **which parts** of an image were likely manipulated and even estimating which AI model might have generated it. It provides an "AI likelihood" score and highlights anomalies, which is a very user-friendly feature for understanding *why* something was flagged. (Illuminarty has a freemium model, with basic detection free and advanced analysis in paid plans.)
- **Microsoft Video Authenticator** – Developed with help from Microsoft's research teams, this tool uses a blend of AI and machine learning to analyze videos and still images for authenticity. It produces a confidence score about whether media has been manipulated and is designed as a scalable solution for businesses and the public. Microsoft released this around 2020, e.g., to help detect deepfakes in election media. It analyzes subtle artifacts (like blending boundaries) in video frames.
- **Intel's FakeCatcher** – A real-time deepfake video detector from Intel. FakeCatcher takes a unique approach by analyzing physiological signals in videos (such as subtle pulse changes in a person's face) that deepfake generators may fail to reproduce. Intel claims this tool can detect deepfake videos with high accuracy and do so *live* (on the fly), which is valuable for content moderation systems.
- **WeVerify (InVID Toolkit)** – WeVerify is a project (backed by the EU) that expanded on the InVID verification toolkit for journalists. It provides tools for verifying images and

videos, including deepfake detection modules. While not a single product, it's a platform that engages experts and the community in collaboratively identifying fake media. Its inclusion in the competitive landscape highlights the *importance of community and open collaboration*, which aligns with Mirage's plan for an open developer SDK and community involvement.

- **Pindrop** – A security company focusing on voice authentication and fraud prevention. Pindrop's technology can detect voice deepfakes (e.g. in phone calls or voice recordings) by analyzing audio characteristics. In the deepfake era, banks and call centers use such solutions to ensure a caller isn't an AI-generated impersonation. (Pindrop was noted among key players in deepfake detection for its audio-focused approach.)
- **Resemble AI** – Known for AI voice generation, Resemble also has capabilities to detect voice cloning or watermark their outputs. They are part of this landscape since voice deepfakes are a growing concern (fraudsters cloning voices of CEOs, etc.). Resemble's dual role in generating and detecting synthesized voice highlights that some companies in generative AI are also providing detection tools to mitigate misuse.

Competitive Differentiation: While there are many players, **few offer a comprehensive multi-modal detector (text, image, video) with a blockchain verification layer**. Most competitors focus on images/videos (and some on audio); text-based AI content detection is a separate field where tools like GPTZero and Turnitin's AI-writing detector exist. Notably, detecting AI-written text reliably remains very challenging – even OpenAI's own text classifier was **discontinued for its low accuracy** (it and other tools like GPTZero “don't work with enough accuracy to rely on” in practice). This suggests an opportunity for *Mirage* if it can innovate in AI text detection (perhaps by using metadata or stylometry alongside ML models). On the blockchain side, a couple of initiatives (e.g. Project Origin, Adobe's Content Authenticity Initiative) use cryptographic **provenance** to verify originals rather than detection – Mirage's approach of storing detection hashes on-chain is relatively novel among current competitors. Overall, *Mirage* will need to compete on accuracy (ensemble models + metadata analysis), speed (real-time responses), and trust (transparency via blockchain, privacy via TEE) to carve out a strong position in this fast-growing market.

Tagline Suggestions

Accompany the name with a concise, impactful tagline. Some **tagline** ideas:

- *"Keep it Real."* – Mira

Each of these can be tweaked to fit the final product name. The tagline should communicate the core value: **detecting fake content and assuring genuineness**, ideally in a tone that

resonates with the intended audience (journalists might prefer a serious trust message, while social media users might appreciate a slightly playful vibe).

Use Cases

There are numerous scenarios where Mirage's deepfake detection agent can be applied. Here are key **use cases** that demonstrate its value:

- **Social Media Fact-Checking Bot:** One immediate use case is the “@agent on any tweet” scenario described. A user on Twitter (X) can mention the Mirage bot under a tweet that contains an image, video, or suspicious text. The bot will analyze the linked content and publicly reply with a deepfake/genuineness score and perhaps a brief explanation. This is powerful for crowdsourced fact-checking – whenever a potentially fake video goes viral, people can summon the bot. It serves casual users and open-source researchers alike. This use case helps build awareness (each bot reply is essentially advertising the service) and directly addresses the main audience of social media users. Over time, such a bot could expand to other platforms (imagine a Reddit bot in comment threads of suspicious videos, or a Telegram bot in group chats).
- **Community Verification Market (PolyMarket-style):** A more novel use case is a **betting or prediction market** around content authenticity. On the Mirage platform, users can submit controversial or viral content (e.g., a video of a political figure that some claim is fake). Other users can then bet “real” or “fake” (yes/no) with cryptocurrency, essentially wagering on the content's authenticity. After a certain period – or after Mirage's analysis plus maybe expert verification – the truth is revealed, and those who bet correctly win a share of the pot while the platform takes a percentage commission. This gamifies deepfake detection and can draw in the crypto community (who enjoy prediction markets). Importantly, it also generates a **valuable labeled dataset** – every piece of content on the market gets scrutinized by users and analyzed by the agent, providing training data and real-world feedback on the model's accuracy. This use case aligns with engaging the community (they're literally invested in the outcome) and provides a revenue stream (via commissions). However, it will require careful design to prevent abuse and to ensure that betting outcomes are resolved by a trusted verification method (combining the AI agent's score with manual fact-checking for high-stakes cases).
- **Journalism and Newsroom Verification:** *Mirage* can be used by journalists, fact-checkers, and media organizations to vet content before publication. In a newsroom scenario, a journalist receives a user-generated video that claims to show something scandalous. Instead of relying on gut feeling or slow manual verification, they run the video through Mirage's agent (via a web dashboard or an API in their content management system). The agent returns a high deepfake probability and highlights inconsistencies in the video's metadata and frames. The journalist then knows to **hold off publishing and investigate further**, potentially preventing a fake story. Conversely,

if the content seems genuine, having a Mirage report as backup adds confidence. Given recent incidents – e.g., AI-generated images causing news scares (like the fake Pentagon explosion image that briefly went viral) – newsrooms are very much in need of such tools. Mirage's transparent approach (with results verifiable on blockchain) is appealing here because the journalist can later show *why* they trusted or doubted a piece of media. This use case taps into a global network of fact-checkers and media outlets that could become early adopters (perhaps through partnerships or trials).

- **Enterprise Content Moderation and Trust & Safety:** Social media platforms and online communities (from big players like Facebook/Twitter to smaller forums and Discord communities) all face the challenge of moderating synthetic media. Mirage can integrate via API to these platforms to automatically scan user-uploaded content. For example, imagine a scenario on a video sharing site: when a user uploads a video, Mirage's API is invoked in the background; if the video is flagged as a likely deepfake or policy-violating synthetic content, it could be queued for human review or even temporarily blocked pending verification. This use case is essentially *B2B SaaS*: providing "Deepfake Detection as a Service" to platforms that want to **minimize misinformation and fake accounts**. Another enterprise angle is in **video conferencing or communication apps** – e.g., a company could use the agent to verify that a person on a video call is real and not a deepfake impostor of an employee (this might sound extreme, but high-security environments are considering such threats). By logging hashes on blockchain, the enterprise even gets an immutable audit trail of content verifications, which could be useful for compliance (imagine a government agency needing to prove later that a particular propaganda video was confirmed fake at a certain time).
- **Financial Services and KYC:** In banking, crypto exchanges, or other fintech, *Know-Your-Customer (KYC)* and identity verification processes are starting to encounter deepfakes. This includes both doctored documents and deepfake **video onboarding** (where someone might use an avatar or synthesized video to fool a live identity check). Mirage can be deployed to detect things like face-swapped video in real time during onboarding, or to scan submitted ID selfies for signs of AI-generation (for instance, some generative artifacts or mismatches between the ID photo and selfie video). The BFSI sector is highlighted as benefiting from deepfake detection to ensure the integrity of digital transactions. Another example is fraud prevention calls: banks have reported cases of voice deepfakes used to trick call center agents. Integrating a voice deepfake detection (maybe via partners like Pindrop or an in-house model) into the call workflow could alert if an incoming caller's voice pattern matches an AI synthesis. The revenue model here would likely be a premium service, but the use case is clear – **protecting financial systems from deepfake fraud**.
- **Government and Law Enforcement:** Law enforcement agencies could use Mirage in forensic analysis of digital evidence. For example, if a suspect presents a video as an alibi or evidence surfaces on social media related to a crime or political event, agencies

need to know if it's authentic. A specialized version of Mirage could be used by digital forensic labs to flag potentially tampered or AI-generated media (as part of chain-of-custody, they could also store the verification proof on-chain for courtroom transparency). Governments are also concerned with deepfakes in political warfare and propaganda. Intelligence units might use the agent to **monitor threat actors** – scanning suspicious videos circulating online that might be state-sponsored deepfakes aimed at sowing discord. This use aligns with the mention that defense agencies see value in such tools to enhance national security. While this is a more long-term, high-end use case (with potentially custom requirements), keeping it in mind can guide the product's robustness (e.g., an "offline mode" of the agent for secure environments, or the ability to run entirely on a private cloud).

- **Content Creators and Platforms (Authenticity Badges):** A positive use case is enabling honest content creators to differentiate themselves. Photographers, influencers, or educators producing genuine content could use Mirage to **verify their own work** and then display a badge of authenticity. For instance, a stock photo website might run every upload through deepfake detectors (including image synthesis detection) and mark content as "Authenticity Verified" if it passes (or mark "AI-Generated" if detected, for transparency). In social media, a creator could voluntarily attach a proof (perhaps via the blockchain certificate feature) that "this video is certified unaltered." This is like a **quality assurance use case** that might become relevant as audiences start demanding proof that what they see is real. It complements the "detect the bad" use case with a "prove the good" feature.

In summary, *Mirage* can be applied wherever digital content's authenticity is in question – which is almost everywhere in today's info-rich world. Initially, focusing on **social media and journalist use cases** will provide visibility and public trust signals, while enterprise integrations (social platforms, fintech, government) will drive revenue once the tech is proven. The variety of use cases also suggests the need for a flexible API and modular architecture to serve everything from a Twitter bot to a closed law enforcement tool.

Revenue Model

Monetizing an AI/web3 product like Mirage can involve multiple streams. Based on the concept's business model brainstorming (SaaS and Verification-as-a-Service), here are the **potential revenue channels** and how they fit the product:

- **SaaS Subscription for Enterprises:** Offer Mirage as a subscription service to organizations that need routine deepfake detection. For example, a media company or social network might pay a monthly or annual fee for a certain number of API calls or content checks. Tiers could be structured by volume (e.g., up to X pieces of content per month) or by features (basic plan: core detection, premium plan: faster responses + detailed reports + on-chain certificates). Enterprises value service level agreements, so

a high-tier subscription might guarantee priority processing, customization options, and dedicated support. This recurring SaaS model would likely be the backbone of revenue once big clients are onboard.

- **Pay-per-Use Verification API (Verification-as-a-Service):** In addition to or instead of fixed subscriptions, a **usage-based** model can cater to a broader range of customers. For instance, a developer or small platform could access the Mirage API and pay per verification (e.g., \$0.001 per image check, \$0.01 per video minute analyzed, etc.). This usage could be metered and billed via crypto (if focusing on web3 clients) or traditional invoicing. The “verification-as-a-service” model means even *individual users* might buy small packages of credits to verify content occasionally (similar to how one might purchase API credits for a few hundred image scans). This lowers the barrier for entry and monetizes at the long tail of many small users.
- **Freemium Model to Attract Users:** While not a direct revenue source, having a free tier or free tool (like the Twitter bot or a limited web demo) is a strategy to build a user base. Over time, some percentage can be converted to paid plans. For example, Mirage might allow 10 free checks per day for any user, but require a paid plan or token stake for heavier use. The free usage serves marketing purposes and data collection (more content to improve models). Monetization comes when users or companies hitting the limits decide to upgrade for higher limits or more features.
- **Prediction Market Commissions:** The second use case described (PolyMarket-like betting on content authenticity) can generate revenue through **transaction fees**. If users are betting on the platform (using crypto or a native token), Mirage can take a small cut of each pot – say 2-5% – for facilitating the market. Platforms like PolyMarket or Augur have used such fee models. Additionally, if Mirage itself participates (for example, seeding markets or matching bets), it could earn from successful outcomes, though that veers away from a neutral position. The more sustainable approach is the straightforward commission on user-generated markets. This revenue stream is somewhat speculative and depends on achieving a critical mass of users who enjoy this feature; however, it could be significant in crypto communities where such prediction markets sometimes handle large volumes.
- **Data and Model Licensing:** Over time, Mirage will accumulate a highly valuable **dataset of deepfakes and genuine content**, plus well-tuned detection models. There could be opportunities to license these models or data to certain partners. For example, a government or large tech company might pay for an on-premise license of Mirage’s detection model (especially if they can’t use a cloud API for security reasons). This is more of a B2B enterprise sales route – custom deals where the model is deployed in the client’s environment (perhaps with periodic updates). Similarly, the curated dataset (minus any sensitive user info) could be licensed to research institutions or used to train other systems, under a paid arrangement. It’s a bit niche but worth considering as the

market matures (since *training data* for deepfake detection is scarce and valuable).

- **Token Economy & Fees:** If a **native token** is introduced for the Mirage ecosystem (more on utility in the next section), revenue could also come indirectly from the token model. For instance:
 - Charging a small token fee for each on-chain verification record (like a gas or service fee when writing the hash to NEAR) – these fees could accumulate in a treasury.
 - If users stake tokens for using the service (e.g., stake X tokens to get unlimited basic access), the team could allocate a portion of those tokens as revenue (or use them to fund operations by selling on the market in a regulated way).
 - Token-based *premium features* – for example, holding a certain amount of token might grant access to advanced features; this can drive token demand which financially benefits the project (though not “revenue” in a traditional sense, it can increase the value of the project’s own token holdings or treasury).
- **Professional Services & Custom Integration:** In the early stages, a practical revenue source might be offering **consulting or custom development** for pilot partners. For example, a news organization might pay to have Mirage customized to their workflow (integrated with their CMS, training the model on their content). Or a government might fund a proof-of-concept deployment focused on a specific threat (like election deepfake monitoring). These one-time or project-based revenues not only bring in funds but also deepen relationships with key clients and generate case studies. While not infinitely scalable, they can be important for a young startup (common in B2B SaaS, where initial custom projects lead to generalized products).

Overall, Mirage’s revenue model is likely a **hybrid of SaaS and transaction-based income**. The core stable revenue will come from offering the detection service (via API or platform) to those who need it regularly (enterprises, platforms) – this is predictable and can grow with the client base. The more innovative crypto-native revenues (prediction market, token fees) can add upside if the platform gains traction among a large user community. It will be important to balance the *community utility* (keeping some features free or affordable to drive adoption) with the *paying offerings* (where heavy users and businesses support the project financially). Given the sensitivity around deepfakes, demonstrating value (possibly via a free tier and success stories) will be key to convincing organizations to allocate budget for this service.

Roadmap

To execute this vision, a clear roadmap with milestones is essential. Below is a phased **roadmap** from the current post-hackathon stage toward a full-fledged product:

- **Immediate Term (0–3 months) – MVP Development:** After the hackathon win, focus on turning the prototype into a robust **MVP**. This includes refining the deepfake detection models (improving accuracy of the DinoV2 + custom model ensemble, testing on more data), and hardening the integration between the Phala TEE agent and the NEAR blockchain logging. Build a simple frontend or dashboard demonstrating end-to-end functionality: e.g., a web interface where a user can upload a piece of content and get a result with an on-chain transaction ID for the hash. Simultaneously, deploy the **Twitter bot** for limited testing on X (perhaps in private or in a small community) to gather feedback. At this stage, also prioritize documentation: an installation guide, architecture diagram (the “arc.png” system architecture should be finalized), and developer docs for the open SDK. Essentially, lay the technical foundation and ensure the core pipeline (frontend → TEE agent → detection model → blockchain log → result) works reliably in a controlled environment.
- **Short Term (3–6 months) – Beta Launch and User Feedback:** Open up the MVP to a wider audience as a **beta**. This could involve launching on NEAR’s testnet or mainnet (depending on confidence) and allowing early adopters (especially in the NEAR community and hackathon circles) to try the agent. During this period, implement **metadata forensics** features (EXIF analysis, video encoding pattern checks) into the agent, so that it’s not solely ML-driven – this can improve accuracy on simpler fakes. Start integrating **additional models** if available (for example, if new state-of-the-art deepfake detection models or Google’s open-source SynthID image detector become accessible, incorporate them). Collect usage data and feedback: which types of content are giving false negatives/positives? Use this to guide model improvement. On the business side, this is a good time to refine the **business strategy** – for instance, finalize pricing models, identify 1–2 pilot enterprise customers or design partners (maybe via NEAR or personal networks) to test the SaaS offering in a real workflow. Also, initiate legal and security reviews, particularly if considering a token or handling user data, to ensure compliance and trustworthiness from the start.
- **Mid Term (6–12 months) – Feature Expansion and Partnerships:** At around the one-year mark, aim to move from beta to a **full launch**. By now, the core detection accuracy should be high and validated by beta users. Introduce **multi-platform integrations**: for example, release the browser extension for Chrome/Firefox that interacts with the Mirage service, and perhaps a plugin for common messaging apps or content management systems (WordPress plugin for bloggers to check images, etc.). Launch the **community verification market (prediction market)** feature in a controlled way – maybe as a pilot event around a specific viral video to test engagement and the mechanics. If a token is to be used, this period might involve creating the token on NEAR, testing its utility in the platform (for betting or staking), and auditing the smart contracts. Parallely, form partnerships: for instance, an alliance with a fact-checking

organization (they use Mirage and provide feedback/data, Mirage gives them free access or co-branded tools), or a collaboration with NEAR (if not already) to be showcased as a prime example of AI x blockchain utility. Marketing efforts should ramp up here too – attending industry events (AI conferences, journalism tech conferences) and showcasing the product. Also, consider **scaling the team**: bringing on additional AI researchers (to keep up with new deepfake techniques), a community manager (to handle outreach and support the user base), and business development roles (to convert pilot users into paying customers).

- **Long Term (12–24 months) – *Scaling and Monetization*:** In this phase, focus on scaling usage and revenue. By 12-18 months, the goal is to have several **paying enterprise customers** (social media platforms, media companies, etc.) and a growing base of active users (including free-tier users driving community). Work on optimizing the **performance** of the service – for instance, reducing the analysis time for videos, improving the throughput if thousands of requests come in (maybe employing a distributed network of TEE nodes to handle load). This is also the time to **implement advanced token economics** if going that route: e.g., introduce **staking mechanisms** (community members stake tokens to get certain privileges or to earn rewards for providing computing resources if a decentralized detection network is in vision). Expand the platform's reach by adding support for more file types or new deepfake modalities (for example, deepfake **audio-only clips**, or even detection of synthetic **text transcripts** in multiple languages as mentioned). Geographical expansion can take place – ensure the UI and documentation are available in key languages (Chinese, Spanish, etc.) to encourage global adoption, especially since deepfake issues are worldwide. If not already done, pursue certification or third-party **evaluations of the technology** (for example, participating in NIST's media forensics challenge or getting academic papers out) to solidify credibility. In this period, the project might also consider **open-sourcing certain components** to build trust and community contribution – perhaps the SDK or older versions of the model, while keeping the latest proprietary for commercial edge.
- **Beyond 2 Years (24+ months) – *Evolution and Industry Leadership*:** The goal is to position Mirage as a **leader in the deepfake defense space**. In the longer term, the roadmap includes contributing to or setting industry standards for content authenticity. Mirage could integrate with initiatives like the **Coalition for Content Provenance and Authenticity (C2PA)** (which Adobe, Microsoft, BBC and others are driving) to ensure compatibility with any emerging standard for tagging real vs fake content. The platform might evolve into a **decentralized network** for content verification – for instance, allowing independent nodes (maybe run by universities or companies) to host the detection models within TEEs and form a *distributed verification network*, with blockchain as the coordinator (this ties back to token incentives if each node earns tokens for processing). By this time, deepfake generation technology will be even more advanced, so a lot of R&D must continue – possibly using *generative AI* itself to come up with detection strategies (an arms race scenario). The roadmap should plan for a dedicated R&D wing to keep improving the AI, possibly in collaboration with academic researchers.

In terms of features, perhaps introduce a **reputation system** for content sources: using the data of verified genuine vs fakes, Mirage might score content publishers (e.g., a social media account that consistently posts deepfakes gets a low trust score) – this can be offered to platforms to integrate as a signal. And finally, consider expanding the use of the technology beyond just detection: for example, **preventive tools** (like an AI plugin for creators to self-watermark their content so it's easier to verify) – this closes the loop by not just catching fakes but helping promote truth from the source.

Each stage of the roadmap should be revisited based on feedback and market developments. Flexibility is key, as the deepfake landscape and competitive scene can change rapidly. However, this phased approach ensures that early efforts (MVP, pilots) build the foundation for later scaling (enterprise adoption, community network effects) and eventual industry influence.

Potential Token Utility

If Mirage introduces a **crypto token** as part of its ecosystem, it will be important to imbue it with meaningful utility. A token can enhance the platform's decentralization and community engagement. Here are several potential uses and utilities for a Mirage token:

- **Access and Payments:** The token could act as a **payment method** for the service. Users (especially in the crypto community) might use the token to pay for verification requests or subscriptions, perhaps at a discounted rate compared to fiat. For example, 100 tokens might grant a month of premium access or a bundle of content checks. This would drive demand for the token as the user base grows. Alternatively, holding a certain number of tokens might automatically give access to premium features (“hold-to-use” model), which can encourage broader distribution of tokens.
- **Staking for Verification Nodes:** In a more decentralized future, if community or third parties run detection nodes (e.g., operating the Mirage agent in a TEE on their own hardware to contribute processing power), those node operators might need to **stake tokens** as collateral. Staking can ensure good behavior – if a node tries to cheat or output false results, they could lose their stake. Meanwhile, honest node operators earn token rewards for their work (kind of like miners/validators in blockchain). This model can help **scale the system** (more nodes = more parallel analyses) and align incentives for accuracy and uptime. Even outside of running nodes, tokens could be staked by users or experts who want to vouch for certain content; for instance, an expert fact-checker stakes tokens on a verification outcome, increasing confidence (and earning a reward if they are correct and helpful).
- **Governance and Voting:** A classic utility for crypto tokens is **governance**. Mirage could become a DAO-like community where token holders can vote on proposals – such as which new feature to prioritize, which partnerships to pursue, or changes to the fee structure. For example, if there's a proposal to adjust the commission rate on the betting

market, token holders could vote on it. Or if adding a new detection model (perhaps one contributed by the community) is on the table, the community might vote to allocate treasury funds to that effort. This turns users into stakeholders in the project's direction. Given the community-oriented branding, this would reinforce that ethos by giving the community a real voice. Care would be needed to prevent governance attacks (hence possibly requiring some staking or identity checks for voters).

- **Rewards and Incentives:** The token can fuel a **rewards system** to encourage beneficial actions:
 - **Reporting Bounties:** If a user discovers a deepfake that the system missed or finds a novel type of deepfake, they could be rewarded in tokens for submitting it (helping improve the model).
 - **Bug Bounties:** Technical contributors who find vulnerabilities (especially important in a security-focused project with TEE and blockchain) might earn tokens.
 - **Prediction Market Rewards:** In the Polymarket use case, tokens could be used for betting, but also the platform might periodically airdrop small token rewards to active participants or to those who consistently vote with the majority (i.e., are good at spotting fakes).
 - **Content Creators Incentives:** One could imagine a system where content creators who consistently produce verified genuine content get token rewards or badges – this might encourage positive behavior (though this is an area to be careful so as not to unintentionally reward simply producing non-AI content over creative AI usage that's properly disclosed).
- **Premium Features via Token Holding:** Similar to how some platforms have NFT or token gating, Mirage could have a model where **holding a certain NFT or amount of tokens unlocks special capabilities**. For example, an NFT that represents a "Pro Verifier Pass" could be bought with tokens and gives the holder access to unlimited personal-use verifications or advanced analytics tools. This kind of utility can drive a mini-economy where power users or enterprises might accumulate tokens/NFTs to ensure they have the best access, and those tokens ultimately flow into the project's treasury or are burned (reducing supply).
- **Burn Mechanism for Sustainability:** To counterbalance token rewards, Mirage could implement a **burn mechanism** where a portion of fees is burned (removed from circulation). For instance, every time someone writes a verification hash to the blockchain, maybe a tiny amount of token is burned as a proof-of-use. This could make the token deflationary with platform growth, which might increase its value and therefore the project's market cap, indirectly benefiting funding. This has to be tuned carefully to

not overly tax users, but it's a common approach in crypto tokenomics to create a feedback loop between usage and token value.

- **Marketplace Currency:** If the platform evolves to host a **marketplace** (say, for deepfake detection plugins or for exchanging datasets, or even the betting market itself), the token can serve as the primary currency within that ecosystem. This ensures an internal economy – for example, a developer could create a specialized model (maybe one that detects deepfake voices in a certain language) and offer its use for a fee in tokens; users pay in the token to use this add-on model and the developer earns tokens. This scenario leverages the open SDK and community contributions: the token helps facilitate a **marketplace of extensions or data** that enrich the main service.

In implementing a token, it's crucial to also consider regulatory and user experience aspects. Not all users (journalists, enterprises) will be comfortable dealing with tokens or crypto wallets. One approach is to keep token usage mostly in the background or optional – for instance, enterprise clients can still pay in fiat and the platform can internally handle tokens for them if needed for on-chain actions. The token should **add value** to the ecosystem (through decentralization, engagement, or network effects) rather than just be a fundraising mechanism. By aligning the token design with the platform's core goals (accuracy, trust, community participation), Mirage can create a sustainable token economy that incentivizes the right behaviors (honest verification, continuous improvement, widespread adoption) and provides an additional layer of product stickiness.

Go-To-Market (GTM) Strategy

Launching and growing Mirage will require a well-planned go-to-market strategy that targets the right users and builds credibility. Below is a multi-faceted GTM approach:

- **Leverage the Crypto/Web3 Community Early:** Since Mirage is built on NEAR and emerged from a Web3 hackathon (winning 1st place in the Oz Builder Day AI x Web3 Hackathon), the NEAR and broader crypto community is a natural first adopter group. Start by promoting Mirage within NEAR forums, Discords, and social media. Write a case study or blog post about the hackathon win and share it on platforms like *NEAR's official blog*, *Dev.to*, *Medium*, etc., highlighting the innovative use of blockchain and TEE in deepfake detection. Crypto users, especially those active on Twitter and in tech circles, are often early tech adopters and will appreciate the blockchain angle. This community can provide initial users for the Twitter bot (they'll test it on content for curiosity) and potentially early token holders if a token is launched. Also, consider cross-promotion with other projects in the NEAR ecosystem – for example, if there's a NEAR-based social media dApp or a content platform, integrate Mirage there and announce the partnership to both user bases. The key here is to gather a **tech-savvy user base** that provides feedback and evangelizes the product within their networks.

- Showcase on Social Media via the Bot:** The Twitter (X) bot itself is not just a feature, but a marketing vehicle. Ensure that when it replies, it does so in a way that piques interest (e.g., including a brief “Analyzed by Mirage AI – Authenticity Score: 95% genuine” and maybe a link to learn more). If the bot publicly flags a high-profile fake (say a celebrity deepfake video or a fake news image) and that gets noticed, it could generate a lot of viral attention. To boost this, the team can proactively use the bot on trending content. For instance, when the next big questionable image or video surfaces (they happen almost weekly somewhere, e.g., a fake quote image or a purported leaked video), have the bot or official account respond quickly with the analysis. This can catch journalists’ eyes. Engaging with misinformation watchdog accounts on Twitter (like those that debunk fake news) by providing them an easy tool will also help. Over time, aim to get the bot “verified” or recognized so users trust it. Also, expand similar bots to other platforms popular with the target audience: a **Reddit bot** could be huge in subreddits like r/deepfakes or r/technology, automatically commenting when users post content asking “is this fake?”. Even a **Telegram bot** for group chats could be offered. The strategy is to meet users where suspect content is being shared, and make Mirage a known helper in those contexts.
- Strategic Partnerships (Journalistic and Platform):** Forge partnerships with organizations that lend credibility and offer reach. For journalists: approach international fact-checking groups (e.g., IFCN – International Fact-Checking Network, or groups like PolitiFact, AFP Fact Check, etc.) and offer Mirage for free in a trial. If those organizations find it useful and mention it in their reports or social media, it’s a huge trust signal to others. Also, seek collaboration with the **Trusted News Initiative** or the Content Authenticity Initiative – these are networks that include major media and tech companies focusing on solutions to misinformation. Being part of such initiatives can open doors (for example, the BBC or Reuters might want to test your tool). For platforms: identify at least one or two mid-sized social networks or community platforms (could be a crypto-native one, like a decentralized social media on Web3, or a traditional forum platform) and integrate Mirage at a pilot level. For instance, if there’s a popular Discord community (say a large crypto Discord) suffering from lots of fake account videos, offer them a bot or API access to moderate content – in exchange, perhaps they’ll officially announce using Mirage. Each successful partnership can be turned into a case study or press release (e.g., “XYZ News Agency uses Mirage to vet incoming footage”), building momentum.
- Content Marketing and Thought Leadership:** Position the Mirage team as thought leaders in the deepfake mitigation space. This involves producing **valuable content**: regular blog posts or reports about deepfake trends, statistics, and notable incidents. For example, publish a quarterly “Deepfake Threat Report” that aggregates info about how many deepfakes were detected, new deepfake techniques seen in the wild, etc. Sensity AI has done similar reports to great effect, getting cited in news articles for stats on deepfake growth. If Mirage can provide fresh data (especially leveraging its platform’s analytics globally), journalists will quote these reports, indirectly advertising the solution.

Additionally, speak at webinars, Twitter Spaces, and conferences about the intersection of AI and blockchain for truth verification. The unique angle of using NEAR blockchain and TEE is likely to attract invitations to panels or podcasts, as it's a novel approach. By raising the profile of the problem (deepfakes) and presenting Mirage as a credible solution, you create pull demand (organizations will come to you asking to try the product they heard about).

- **Community Building and Developer Outreach:** Since an open developer SDK is part of the plan, actively cultivate a **developer community**. Host hackathon-style events or challenges where developers are encouraged to build something using Mirage's API/SDK – e.g., “Who can build the coolest app that uses Mirage to fight deepfakes?” Offer prizes (could be tokens, could be cash or just recognition) for winners. This can bring in innovative integrations: someone might build an AR app that uses Mirage to check if what you're seeing through your phone camera is likely real, or a plugin for video editing software to self-check content. By encouraging these projects, Mirage extends its reach organically. Make sure to have a presence on developer platforms: a well-documented API on GitHub, a Postman collection for the API, and a support channel (like a Discord or Stack Overflow tag) for technical questions. Every developer who integrates Mirage brings along their user base as potential indirect users of Mirage. Additionally, involve the community in improving the core product – for example, start a **community deepfake database** where volunteers can contribute examples of fakes they find, and perhaps label them, with top contributors acknowledged or rewarded. This not only helps the model but makes the community feel they are part of the mission “to safeguard digital media integrity.”
- **Freemium Conversion Strategy:** Use the free tools (bot, limited web checks) to funnel users into paid offerings. On the website or app, implement gentle upsells: after a user has used 5 free verifications in a day, prompt them: “Like using Mirage? Pro plan gives you unlimited checks and detailed reports.” Ensure the sign-up and payment process is smooth (if targeting crypto users, allow payment in crypto; for others, provide credit card or even a fiat subscription with invoice for enterprises). Possibly offer a **free trial** of premium features for journalists or enterprise leads to hook them (for example, “Journalists get 1 month free of our Pro tier – sign up with your news org email”). Also, consider a referral program: existing users (especially community members) who bring in new users could get bonus credits or tokens. This encourages word-of-mouth growth.
- **PR and Media Outreach:** Capitalize on the hackathon victory and any early successes by proactively engaging with tech media. Draft press releases for significant milestones (e.g., product launch, partnerships, notable funding or grants, major new feature). Send these to journalists covering AI, blockchain, or cybersecurity. Many media outlets find the deepfake topic hot and the twist of using blockchain interesting – headlines like “Startup Uses Blockchain and AI to Expose Deepfakes” could garner clicks. Offer to demo the product to journalists – perhaps analyzing a deepfake of their choice live – this could turn into articles or news segments. Also, be prepared for crisis-response PR: if a huge

deepfake story breaks (like a political deepfake), quickly publish commentary or have the team available to speak as experts. Being quoted as *experts* in news stories about deepfakes will indirectly validate Mirage's expertise. Over time, aim for one or two marquee media coverages (for instance, an interview in Forbes or TechCrunch about Mirage's mission and tech).

- **Targeted Outreach to High-Risk Sectors:** As the product stabilizes, do direct outreach to entities that *must* care about deepfakes. For example, approach **election commissions or political parties** ahead of elections to offer your services (perhaps at a discount initially) to monitor for fake videos targeting their candidates. Similarly, approach large fintech or banks' fraud departments about the rising threat of voice/video deepfakes in scams – offer a pilot integration of Mirage in their process. These are more traditional sales efforts (whitepapers, demos, meetings) but given the novelty of deepfake threats, some may be very interested in a solution. Emphasize how Mirage's multi-layered approach (ML + metadata + blockchain audit trail) can give them an edge in security and compliance. Success in even one such high-risk sector (say, a bank adopting it, or an election monitoring use) can be publicized to accelerate adoption in other similar entities.
- **Iterate and Localize GTM:** Gather metrics on what channels are bringing the most engagement. Maybe the Twitter bot goes viral and becomes the main entry point – then invest more in it (maybe multiple language bots?). If developers show big interest, maybe host more hackathons. Stay agile; the deepfake detection market is new, so part of GTM is discovering where the resonance is strongest. Also, plan to **localize** efforts: for instance, deepfake problems are very pronounced in certain regions (e.g., misinformation in India or Brazil on WhatsApp). If targeting globally, consider local ambassadors or content in different languages. A community-oriented strategy could involve identifying champions in different countries (e.g., a journalist or professor who loves the tool) and equipping them to spread the word locally – possibly via a formal ambassador program or just informal support.

Success in the GTM phase will be measured by growing usage (both free and paid), the quality of partnerships established, and the reputation Mirage builds as a go-to solution for deepfake detection. By combining grassroots community building (typical in Web3) with B2B relationship building (needed for enterprise adoption) and a strong public narrative (thought leadership in fighting misinformation), Mirage can accelerate from a hackathon prototype to a widely recognized platform. The problem it tackles is timely and widely acknowledged; the strategy is about positioning Mirage as an *innovative, trusted, and accessible* answer to that problem, and doing so before others dominate the mindshare in this emerging market. With the above approach, Mirage should be well on its way to securing significant market share in the deepfake detection landscape while staying true to its community-driven values.

Current MVP Benchmarks

Image Deepfake Detection

Model / Approach	Dataset / Task	Metric	Reported Performance	Notes / Comments
Mirage	DFDC (DeepFake Detection Challenge) and FaceForensics++	Accuracy	~97%	Transformer-based architecture with top performance on DFDC and FaceForensics++. It uses transfer learning on large-scale datasets for high generalizability.
VGG16 + LSTM	FaceForensics++	Accuracy	~91%	Combines VGG16 for feature extraction and LSTM for sequence learning. Good for identifying deepfake artifacts in both spatial and temporal dimensions.
ResNet-50 + GANs (Generative Adversarial Networks)	Celeb-DF	Accuracy	~98%	Uses ResNet-50 for detecting facial manipulation, augmented by GANs to detect high-frequency artifacts that are indicative of deepfakes.
Face X-ray	FaceForensics++	AUC (Area Under Curve)	~99.5% AUC	Focuses on detecting forensic artifacts that result from deepfake generation (e.g., poor lighting, texture anomalies). Very robust against GAN-based deepfakes.
MesoNet	FaceForensics++	Accuracy	~95%	Designed specifically for fake face detection using mesoscopic patterns in face images. Performs well on medium-quality deepfakes.

DFDNet (DeepFake Detection Network)	DFD (DeepFake Detection)	Accuracy	~98%	Focused on synthesizing fake faces, DFDNet uses contextual features and inconsistencies for detection. Very effective at detecting low-quality deepfakes that appear more natural to the human eye.
Xception + LSTM (Temporal Feature Analysis)	UADFV (Unrestricted ADvanced Fake Videos)	Accuracy	~94%	This architecture combines the Xception network with LSTM to account for temporal dynamics in deepfake videos. It excels at detecting inconsistencies across frames.
EfficientNet + GANs	FakeAVCeleb	Accuracy	~97%	EfficientNet combined with GANs for fine-grained texture and spatial feature extraction to detect deepfake videos at high resolution.
TecoGAN	Celeb-DF, DeepFake Detection Challenge	Accuracy	~95%	A video-based deepfake detection model that uses Temporal Consistency across frames to identify irregularities in motion and appearance, particularly in high-quality deepfakes.
Capsule Networks (CapsNet)	DeepFake Detection Challenge (DFDC)	Accuracy	~93%	Capsule Networks have shown potential in learning spatial relationships in images, which helps detect deepfake artifacts that would typically be missed by traditional CNNs.

Dual-Stream CNN	FaceForensics++	Accuracy	~97%	This architecture separates temporal and spatial features, which enhances its performance in detecting deepfakes involving facial manipulations across videos.
-----------------	-----------------	----------	------	--

Text Deepfake Detection

Model / Approach	Dataset / Task	Metric	Reported Performance	Notes / Comments
Mirage	AI-generated text vs Human text	Accuracy	~93–96%	Fine-tuned transformer model designed for deepfake detection, specifically optimized for identifying inconsistencies in AI-generated content like news articles, research papers, and general conversation
RoBERTa-based Approach (Fine-tuned)	Mixed human and AI-generated texts	Accuracy	90–96%	Fine-tuning of transformers like RoBERTa can be effective for detecting syntactic and stylistic anomalies in AI-generated text
DetectGPT (Fine-tuned GPT-2 model)	Human vs. GPT-2 generated text	Accuracy	70–80% (on average)	Focuses on detecting artifacts and the unique characteristics of GPT-2 generated text. Performance degrades with paraphrasing
Turing-NLG-based Model	Synthetic vs. Human text (mix of models)	Accuracy	93% (best reported)	Uses large-scale, pre-trained models like Microsoft's Turing-NLG to differentiate between text generated by AI models and human-written text.

				Performs well with large datasets, but struggles with paraphrased AI text
GLTR (Giant Language Model Test Room)	GPT-2, GPT-3 output detection	Visual / Statistical Features	Detects statistical patterns in text (probability, word choice predictability)	Good for spotting “unnatural” predictability in AI-generated text. (Not a standalone classifier but a helpful tool for analyzing likelihood distributions in generated text.)
BERT-based Classifier	User-generated content from social platforms	F1 Score	~85–90%	Fine-tuned for detecting AI-authored posts by focusing on linguistic anomalies, style differences, and inconsistency in textual features.
GPTZero (Custom GPT-3 Detector)	GPT-3 generated text vs human-written	Accuracy / F1	91%+	Specially designed to detect GPT-3’s specific patterns, including content generation style and sentence-level coherence. It’s particularly good with essays and long text pieces
Turing Test for Text Generation	General AI vs. human texts	Accuracy	85–95%	This test, along with automated tools, evaluates whether AI-generated text can pass as human-written. While not a model per se, it’s an important reference for baseline evaluation.
Transformers-based (Longformer)	Long-form AI-generated text (research papers, articles)	Accuracy	90%	Longformer, adapted for analyzing long-form content, performs well in detecting long, coherent AI-generated articles where traditional transformers struggle due to input

				length limitations
DistilBERT for Text Classification	General AI vs Human text (research, blogs, forums)	F1 Score	90–94%	Light-weight BERT model that performs well in settings where speed and real-time processing are required (e.g., moderation of online forums). Accuracy holds for a wide range of datasets
LIAR-PLUS (with fine-tuned models)	Fake news detection	Accuracy / F1	~91%	A specific model aimed at detecting fake news by evaluating the authenticity of news articles. This approach also applies to text deepfake detection when a dataset of AI-generated news content is used

Audio Deepfake Detection

WIP

Video Deepfake Detection

WIP

Links to MVP

<https://askmira.io/>

<https://miragetech.io/>