WITNESS' TRIED Benchmark: A Checklist for Truly Innovative and Effective AI Detection

Instructions

The checklist below is adapted from the benchmark quality assessment framework outlined in *BetterBench* by Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, and Mykel Kochenderfer from the Stanford Intelligent Systems Laboratory. Inspired by their methodology, this Benchmark uses a similar checklist format, distributes considerations alongside the stages of the tool's lifecycle and introduces point scale reflecting different phases of realization of set criteria.

The checklist is designed to guide AI detection developers in evaluating whether their tools align with best practices for equitable and effective detection throughout the stages of **design, development, testing, implementation, and maintenance**. Each question in the checklist should be answered with a Yes, No, TO DO or N/A, followed by a concise justification (approximately one sentence). Justifications may reference specific page numbers from your paper or include relevant links to additional information.

Each answer receives a number of points. The points are awarded in the following manner:

Answer	Number of points
Yes	3 points
To do with justification	2 points
To do without justification	0 points
No with justification	1 point
No without justification	0 points

Upon completing the checklist, you will have the opportunity to calculate your score manually. At the end, you will be provided with a minimum score to evaluate the overall equitable effectiveness of the tool. These scores serve as a guide to help you assess the quality and rigorousness of your tool's development process.

This checklist accompanies the report ***TRIED: Truly Innovative and Effective AI Detection Benchmark, developed by WITNESS*** which can be accessed on Arxiv.

Benchmark Design

1. The goals, key concepts, primary features, and target audience for the detection tool are clearly outlined.
2. The design process actively involves input from domain experts with relevant expertise.
3. Relevant academic research, industry standards, and existing literature are thoroughly reviewed and integrated into the design.
4. Real-world scenarios and practical use cases are incorporated to guide the tool's development and application.
5. The mechanisms to ensure the tool's durability and adaptability to rapid development of synthetic media are defined and prioritized in the design process.
6. The tool's funding allows for a responsible and sustainable development and operation of the tool.

Design checklist

1. The goals, concepts, characteristics, and audience of the detection tool are defined.

   1.1. The intended audience is clearly defined.

   ☐ TO DO ☐ YES ☐NO ☐ N/A

   Justification:

   1.2. The use cases on which the tool should be used are clearly described.

   ☐ TO DO ☐ YES ☐NO ☐ N/A

   Justification:

   1.3. The goals and the way the detection model works and is designed are explicitly communicated.

   ☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

1.4. Differences between this tool and existing detection tools are outlined, including unique capabilities or improvements.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

2. The design process involved consultation with diverse domain experts, including those from global and underserved contexts.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

3. Diverse relevant academic research, industry reports, and existing literature were integrated into the tool's design.

3.1. The research and literature were diverse and steps were taken to include knowledge produced outside of Europe and the United States.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

4. Real-world scenarios and practical use cases are incorporated to guide the tool's development and application.

4.1. Real-life cases and use cases were incorporated to reflect practical challenges and expectations.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

4.2. Stakeholders representative of the target audience were consulted in the design phase to clearly outline their key needs and expectations.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

5. The mechanisms to ensure the tool's durability and adaptability to rapid development of synthetic media are defined.

5.1. Specific steps are outlined to ensure that the tool is future-proof and guarantee that it will remain relevant in the long term.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

5.2. The tool's funding allows for a responsible and sustainable development and operation of the tool.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

The maximum number of points is 30.

Your score is:

Benchmark Development

1. The development stage requires explicit consideration of underrepresented languages, cultural nuances, and geopolitical challenges in both training and testing.
2. Training data is diverse, representative of global demographics, and collected ethically, with transparent documentation of the sourcing process.
3. Comprehensive and accessible documentation is provided, integrating relevant context to aid understanding and usability.
4. The tool's limitations are clearly identified and communicated to ensure realistic expectations of its capabilities.
5. Mechanisms for explainability are incorporated, ensuring that results and detections are interpretable by both technical and non-technical users.
6. The development team is diverse, reflecting the perspectives and needs of the tool's intended audience.
7. Accessibility is prioritized, ensuring the tool is usable by the intended audience regardless of technical expertise or resource constraints.
8. The tool's capabilities for handling various file types and quality levels are explicitly detailed.
9. The tool demonstrates consistent performance across diverse contexts and use cases it aims to serve.
10. Development respects and integrates with existing verification techniques and skill sets to enhance reliability and usability.
11. Adaptability is a key focus, allowing the tool to evolve in response to new challenges, use cases, and technological advancements.

Development Checklist

6. Training data is diverse, representative of global demographics, and ethically sourced.
   6.1. Data collection process complied with the local data protection regulations.

   ☐ TO DO ☐ YES ☐ NO ☐ N/A

   Justification:

   6.2. The training dataset is representative of diverse demographics.

   ☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

6.3. Developers implemented measures to assess and mitigate biases during training.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

7. The tool's capabilities for handling various file types and quality levels are explicitly detailed.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

8. The training dataset included examples of corrupted synthetic content.

8.1. The tool was trained on low-resolution content.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

8.2. The tool was trained on social media compression standards.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

8.3. The tool was trained on reformatted content, including files resaved in a different format.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

8.4. If the tool is capable of detecting audio, it was trained on audio content corrupted by noise (including background noise, music and cross-talking).

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

8.5. If the tool is capable of detecting video, it was trained on video content with dynamic movements and noisy background.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

9. The detection tool was trained to deal with diverse types of content.

9.1. If the tool is capable of detecting audio, the training dataset included data with a transmission similar to telephones.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

9.2. If the tool is capable of detecting audio, the training dataset included radio conversations (such as radio broadcasts and walkie-talkie conversations).

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

9.3. If the tool is capable of detecting video, the training dataset included footage featuring dynamic content of various angles and positions, with movements or multiple people.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

9.4. The tool's training dataset included different types of files, including file compressions on different social media platforms.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

10. The tool is developed to scale across different levels of content complexity, size, and volume.

10.1. A system is in place, and adaptable to adjust the detection model to continually update the training dataset to include new examples of AI-generated or manipulated content.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

11. The tool provides users with clear information on how detection results should be interpreted.

11.1. The tool does not use binary labels (such as ‘real’ or ‘fake’) when communicating the results but instead describes manipulation using accessible description and language of degrees.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

11.2. The tool includes validation processes that go beyond confidence scores, offering meaningful contextual insights or probability distributions.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

11.3. The additional information outlines the tool’s capabilities, including what content it can analyze, top level information on the data it was trained on, and the training objectives.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

11.4. The additional information outlines the tool’s limitations, including the language constraints and the challenges that may impact its performance, such as the quality of analyzed content.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

11.5. The tool points to the specific elements indicating that the content is synthetic, such as the exact moment when the manipulation occurred, specific regions identified as synthetic, or metadata-based insights supporting its conclusion.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

12. The tool acknowledges other existing verification techniques and incorporates information about them into its workflow.

12.1. The information is provided with respect to other verification techniques and skill sets that could be integrated.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

12.2. When the detection tool cannot provide a conclusive result, the accompanying information still provides valuable insights to the user and may be used along with other verification methods.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

13. The tool is adaptable with processes in place to evolve in response to new challenges, use cases, and technological advancements.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

The maximum number of points is 57.

Your score is:

Benchmark Testing

1. Proactive measurements are being taken to test the tool's blind spots.
2. The tool undergoes regular and systematic testing to maintain reliability and effectiveness.
3. Diverse stakeholders and domain experts, including independent external groups, actively participate in the testing process to provide varied perspectives and expertise.
4. The tool is rigorously tested on challenging edge cases, including adversarial content, false claims of AI-generated media, and heavily manipulated content.
5. The tool demonstrates resilience against evasion techniques, maintaining its accuracy and reliability even under deliberate attempts to bypass detection.
6. The tool is evaluated from a human-detection perspective.

Testing Checklist

14. The tool was tested in an exhaustive and inclusive manner.

14.1. The tool is proactively tested to identify blind spots and potential failures.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

14.2. The tool's robustness was tested against adversarial attacks and evasion techniques.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

14.3. External stakeholders were involved in the red-teaming of the tool.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

14.4. The tool was tested by a global group of stakeholders from different regions to identify vulnerabilities and challenges stemming from application of the tool in specific cultural contexts.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

14.5. Testing is conducted to evaluate the tool's performance on real-world examples of deceptive and manipulated content.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

14.6. Results and performance metrics are documented, and top level information is transparent.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

15. The tool is evaluated from a human-assisted detection perspective.

15.1. Such testing included assessing how much time it took to receive the result and how difficult the process was from the human perspective.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

15.2. Such testing included a comparison between the performance of human-assisted detection and detection by an individual not using a detection tool.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

The maximum number of points is 24.

Your score is:

Benchmark Implementation

1. The tool is designed and implemented to uphold human rights, ensuring it does not infringe on privacy, freedom of expression, or other fundamental rights.
2. Accessibility is prioritized, ensuring the tool is usable by its intended audience, regardless of technical expertise or resource availability.
3. Documentation is clear and comprehensive, and the top level version is easily accessible, providing users with the necessary guidance to operate the tool effectively and understand its limitations.

Implementation Checklist

16. Ensure the tool does not inadvertently violate human rights, such as privacy or freedom of expression.

16.1. The tool explicitly avoids outputs or recommendations that could harm individuals or communities, aligning with ethical AI principles.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

16.2. The tool ensures secure handling of sensitive data and complies with relevant data protection regulations.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

16.3. The tool ensures that it does not discriminate in its outputs and that its accuracy does not vary depending on the demographic of the individuals featured in the content.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17. The tool is accessible to a diverse group of targeted users.

17.1. The tool includes educational resources or tutorials to help users understand its functionality and limitations.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17.2. The tool does not require advanced technical knowledge or skills from the user to operate the tool.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17.3. The tool operates efficiently without requiring significant computational power or strong network connectivity.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17.4. The tool has the option to function offline.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17.5. Multilingual support is available for target audiences, and limitations in other language support are clearly communicated.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17.6. The tool is affordable to the targeted group of users.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

17.7. The tool provides clear, coherent summaries of its results tailored to a general audience.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

18. The tool's performance metrics are accurate and not exaggerated.

18.1. The tool communicates errors or uncertainties clearly to users, avoiding overconfidence in ambiguous cases.

☐ TO DO ☐ YES ☐ NO ☐ N/A

Justification:

The maximum number of points is 33.

Your score is:

Benchmark Maintenance

1. The tool undergoes regular performance evaluations to ensure its detection accuracy remains reliable across evolving datasets and content types.
2. Resources are actively being distributed for updates and maintenance processes.
3. Updates are routinely implemented to address emerging challenges, such as adversarial attacks, advancements in deepfake technologies, and new forms of synthetic content.
4. User feedback is actively collected, documented, and incorporated into updates, with a transparent and accessible communication channel available for issue reporting and suggestions.
5. Clear policies are established regarding the support duration for older versions of the tool following updates or new releases.
6. Periodic audits are conducted to identify and mitigate any biases introduced during updates or revealed by newer datasets.
7. Developers continuously monitor advancements in AI and related technologies, ensuring the tool remains aligned with the latest innovations and best practices.

Maintenance Checklist

19. The tool undergoes regular updates, and changes are documented transparently.

19.1. The tool's accuracy is evaluated against previously verified cases.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

19.2. Clear policies are in place for how regularly updates will be conducted.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

19.3. The tool is periodically reviewed for its applicability and usability across diverse cultural and linguistic contexts.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

19.4. Regular audits are conducted to detect and address bias introduced in updates or uncovered in newer datasets.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

19.5. The tool communicates to the user when it was last updated.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

19.6. Documentation is kept up to date with every tool revision, including new use cases, limitations, and changes in features.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

20. An accessible channel of communication is provided.

20.1. Feedback channels for users are actively maintained, and updates incorporate user feedback.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

20.2. A contact person or team for support and inquiries is clearly listed.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

20.3. An active community of users, developers, and domain experts is cultivated to provide ongoing support and collaboration.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

21. The tool has End-of-Life procedures in place.

21.1. The tool will be retired if it has not been updated for a specified time following clear policies.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

21.2. If the tool is to be discontinued, clear communication is provided to users, along with data migration or alternative solutions.

☐ TO DO ☐ YES ☐NO ☐ N/A

Justification:

The maximum number of points is 33.

Your score is:

The total maximum number of points is 177.

A score above 143 points indicates that your tool is **truly effective**.
A score between 107 and 142 points indicates that your tool is **moderately effective**.
A score between 72 and 106 points indicates that your tool is **somewhat effective**.
A score below 71 points indicates that your tool is **not effective**.

Your score: _____