🍩 AI System Card Review Checklist v2

☀️ Sandy Dunn

Who Should Use This Checklist?

- ★ **CISOs** and security architects reviewing AI vendors or APIs
- ★ **AI product teams** evaluating safety before model deployment
- ★ **Risk and compliance teams** conducting due diligence
- ★ **Procurement** teams ensuring vendors meet transparency standards

---

System Overview & Fundamentals

- ☐ Clear description of the model's intended purpose and primary function
- ☐ Explicit statement of intended use cases and target users
- ☐ Technical architecture details (model type, size, key components)
- ☐ Identification of unintended use cases or discouraged applications

Data & Training

- ☐ Training data sources, size, and composition documented
- ☐ Information on datasets used for training, validation, and testing
- ☐ Data diversity and potential biases identified and addressed
- ☐ Training methodology and key hyperparameters specified
- ☐ Data privacy and security measures during training and operation
- ☐ Data governance and lineage documentation

Performance & Evaluation

- ☐ Performance benchmarks on relevant tasks with quantitative metrics
- ☐ Evaluation methodologies and datasets clearly described
- ☐ Performance results broken down by relevant subgroups for fairness assessment
- ☐ Comparison to relevant baseline systems
- ☐ Internal testing methodologies and results
- ☐ Performance across different demographic groups and contexts
- ☐ Use of RLHF (Reinforcement Learning from Human Feedback) documented

Capabilities & Limitations

- ☐ Detailed capability assessment with specific tasks the system can perform

- [ ] Known limitations and failure modes explicitly stated
- [ ] Boundary conditions where the model performs poorly
- [ ] Scenarios where the system may underperform
- [ ] Robustness to adversarial attacks assessed
- [ ] Susceptibility to issues like hallucination documented

Risk Assessment & Analysis

- [ ] Comprehensive risk analysis covering potential harms
- [ ] Common risks addressed (bias, misinformation, hallucination)
- [ ] Context-specific risks identified
- [ ] Dual-use potential and misuse scenarios discussed
- [ ] Analysis of societal and ethical implications
- [ ] Environmental impact considerations
- [ ] Privacy and data protection implications
- [ ] Likelihood and severity of risks assessed

Safety & Mitigation Measures

- [ ] Technical safeguards and procedural mitigations for identified risks
- [ ] Information on content filtering and output monitoring systems
- [ ] Information on rate limiting and access controls
- [ ] Information on systems for human oversight and intervention
- [ ] Were Red team testing results included
- [ ] Confirm bias and fairness testing & assessments completed
- [ ] Confirm adversarial testing done

Governance & Compliance

- [ ] Development governance processes documented
- [ ] Confirm compliance with the organizations relevant regulations and standards
- [ ] Information about their version control and update procedures
- [ ] Information about ongoing monitoring and incident response procedures
- [ ] Details on their plan for addressing newly discovered risks
- [ ] Process for ongoing monitoring, updating, and re-evaluation
- [ ] External audits or certifications outlined
- [ ] Includes incident reporting contacts or email and response protocols

Deployment & Usage

- [ ] Includes recommended deployment contexts and restrictions
- [ ] Provide usage policies and restrictions on model use
- [ ] Information on integration requirements and dependencies
- [ ] Their monitoring recommendations for deployed systems
- [ ] Documented user guidance and training materials provided
- [ ] Their support and maintenance commitments

- [ ] Operational and environmental assumptions stated
- [ ] Context for real world deployment described

Transparency & Explainability

- [ ] Information on how system decisions or outputs can be interpreted
- [ ] Contact points for further inquiries provided
- [ ] Alignment or safety objectives listed
- [ ] What is not disclosed and reasons for non-disclosure
- [ ] Disclosures aligned with responsible AI principles

Model Weights & Lineage

- [ ] Provenance & Licensing, verify where model weights are from
- [ ] Are the weights under a license that allows the intended use
- [ ] Confirm there aren’t any downstream restrictions
- [ ] Has the vendor or team modified or fine-tuned the weights
- [ ] Are any changes documented
- [ ] Have the weights been checked for malicious tampering or backdoors
- [ ] Have any changes been tested and evaluated

System Card Review Summary

- [ ] Confirm all critical areas adequately addressed
- [ ] Confirm documentation meets transparency standards
- [ ] Confirm risk mitigation measures are proportionate to identified risks
- [ ] Confirm system appropriate for intended use case
- [ ] Note any red flags or concerns noted for further investigation