Ardigen

Data Usability Checklist

Use this checklist to evaluate if your data are AI-ready. If several of the signals below apply, the dataset may still support exploration, but it is unlikely to support actionable, decision-grade AI.

1. Experimental provenance is unclear

- [ ] You cannot clearly reconstruct how, when, or by whom the data was generated
- [ ] Protocol versions, reagents, instruments, or cell lines are not traceable
- [ ] Raw data is unavailable or irreversibly transformed

2. Metadata is missing, inconsistent, or informal

- [ ] Key variables are stored in free text, lab notes, or file names
- [ ] Different teams record the same parameter in different ways
- [ ] Raw data is unavailable or irreversibly transformed

3. Assay definitions changed over time

- [ ] Assay formats, readouts, or thresholds evolved without explicit versioning
- [ ] Historical and recent measurements are pooled without correction
- [ ] No controls exist to assess assay drift

4. Negative and inconclusive results are absent

- [ ] Only ‘successful’ experiments are retained
- [ ] Failed compounds, weak binders, or null results were filtered out
- [ ] Stopping rules are undocumented

5. The dataset was generated for a single, narrow question

- [ ] Data was never intended to be reused
- [ ] Variables irrelevant to the original hypothesis were not captured
- [ ] No thought was given to future integration or extension

Ardigen

6. Data cannot be linked to decisions or outcomes

- [ ] You cannot trace which decisions were informed by the data
- [ ] Downstream validation results are disconnected or missing
- [ ] There is no feedback loop between prediction and outcome

7 . Integration with other datasets is fragile or manual

- [ ] Merging datasets requires ad hoc scripts or human interpretation
- [ ] Identifiers are inconsistent across systems
- [ ] Ontologies or schemas are undefined or incompatible

8. Dataset size hides structural weakness

- [ ] Large volume but low diversity
- [ ] Many samples, few independent experiments
- [ ] Apparent richness without coverage of edge cases

9. Quality control is implicit, not explicit

- [ ] No documented QC criteria
- [ ] Outliers removed without a rationale
- [ ] Replicates are missing or ignored

10. Reuse assumptions were never discussed

- [ ] No agreement on how the data might be reused
- [ ] Ownership, access, and update rules are unclear
- [ ] The dataset exists, but no one feels responsible for it

How to interpret results

1-2 signals:	Dataset may be salvageable with targeted curation
3-5 signals:	Expect AI results to remain exploratory
6+ signals:	Treat the dataset as scientifically informative but structurally unfit for AI-driven decisions