AI AGENT RELEASE · 32-ITEM CHECKLIST

AI Agent Release-Readiness Checklist

For agentic media workflows. Each item needs an owner and evidence, such as a test report or an eval run.

1. Quality and evals

- [ ] A versioned eval set covers the real content mix and every agent task.
- [ ] Regression and capability evals are tracked separately.
- [ ] Release thresholds are set per task and met on every one of the repeated runs.
- [ ] LLM-as-a-judge graders are calibrated against human reviewers, and the grader model version is pinned.
- [ ] Scores are compared with the previous model and prompt version, per content type.

2. Tool calls

- [ ] Every tool has tests for valid use, invalid parameters, and tool failure.
- [ ] The agent cannot publish, delete, or share without the required confirmation.
- [ ] Publish, export, and billing calls use idempotency keys, and step and retry limits prevent loops.
- [ ] The agent reports failed actions to the user.

3. Performance and cost

- [ ] Time and cost budgets are set per workflow and tested on the longest real uploads.
- [ ] Long jobs survive app backgrounding and network loss on mobile.

4. Security

- [ ] Prompt injection tests cover subtitles, metadata, transcripts, comments, and on-screen text.
- [ ] No content from a media file can trigger a public or destructive action without user confirmation.
- [ ] System prompts and other users’ data cannot be extracted through the agent.

5. Content safety, copyright, and privacy

- [ ] Generated text and images are checked for misleading or off-brand output.
- [ ] Licensing rules apply to any media the agent adds.
- [ ] Personal data does not cross between users, and retention matches the privacy policy.
- [ ] AI-generated media carry the labels and marking required by the EU AI Act.
- [ ] Cloned or synthetic voices are used only with recorded owner consent.

6. Captions and localization

- [ ] Captions are checked for meaning-critical errors, speaker changes, and sound cues.
- [ ] Caption timing stays in sync after agent edits, across the full file length.
- [ ] Each launch language is reviewed by native speakers.

7. Recommendations

- [ ] Relevance and diversity are evaluated for new, narrow-interest, and sensitive-history users.
- [ ] Recommendation explanations do not expose private viewing history.

8. Human control

- [ ] Agent edits are non-destructive and reversible.
- [ ] Users see a summary of changes before accepting them.
- [ ] A named owner can stop the agent, including queued jobs.
- [ ] Prompt and model versions can be rolled back independently.

9. Monitoring and decision

- [ ] Every agent run is traced and searchable.
- [ ] Alerts exist for quality drops, cost increases, and error spikes.
- [ ] A model version change triggers the regression eval before rollout.
- [ ] Every accepted risk has an owner, a mitigation, and a review date.