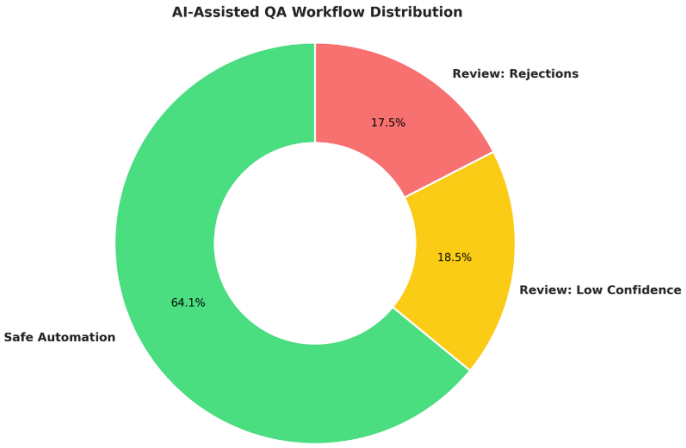
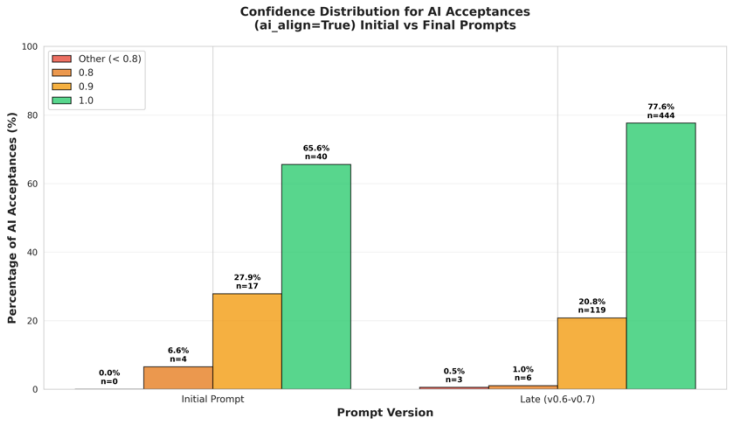
Scaling Video Annotation QA with Agentic AI

When a client requested collection, annotation, and validation of 3,500 videos over 6 weeks, we faced the classic quality-at-scale dilemma. The project required building a high-quality dataset of human-produced question-answer pairs by considering each video's visual content, audio characteristics, and subjective qualities. Contributors would submit videos and answer detailed questions about them, but how could we ensure quality across thousands of submissions? Manual review of every Q&A pair would be costly, and scale 1:1 with project size. Simple automation couldn't handle the multi-modal complexity (video, audio, metadata) and subjective judgments required. A traditional automated QA system would require intensive manual analysis and iterative improvements—time and resources that compound costs in a compressed timeframe. Our solution: an AI agent to 1) automatically validate "clear" Q&A submissions, 2) flag edge-cases for human evaluation, and 3) improve itself by analyzing differences between its evaluations and gold-standard human reviews.



Our evaluation revealed...

- **AI + Human workflow for operational efficiency:** We found that a confidence-based workflow gave outstanding results. Submissions approved with maximum confidence (**~64% of total volume**) are cleared for submission without human review (except for a random sampling to ensure continued quality). All rejections and low-confidence cases (36% combined) route to human experts. This distributes work optimally: AI handles high-volume, clear-cut decisions at **$0.0032 per video**, while humans focus on genuinely difficult edge cases where expertise adds value. **The economics fundamentally change at scale**. As submission volume grows 10x or 100x, AI absorbs the baseline load while human capacity concentrates on high-value judgment calls.



- **The system learns and improves:** Through its self-reflection process, the agent demonstrably improved over time. Initially, 65.6% of AI approvals reached maximum confidence. After iterative prompt refinement guided by the agent's own analysis of where it disagreed with humans, this jumped to **77.6%**, a 12-point improvement in certainty when accepting content.

- **99% precision when the AI is confident:** The AI agent achieved 93.4% overall agreement with human reviewers, but when approving submissions with maximum confidence (1.0), accuracy reached **99%,** nearly perfect precision.