

Table 1 - Reliability Metrics by Model (Audio-only)

Model	Segment Pass Rate	First Try Prob	Two Try Prob	Certified (≤ 2)
Control	50.0%	11.9%	22.4%	No
Base	58.3%	19.8%	35.3%	No
Medium Rubric	66.7%	28.8%	49.6%	No
High Rubric	81.0%	53.4%	78.1%	No
High Rubric + Reasoning	79.8%	51.6%	76.1%	No
High Rubric + Timestamps	79.8%	50.4%	75.7%	No