S2S Models	Response Quality			Content Density		
	T1	*T2*	*T3*	*T1*	*T2*	*T3*
GPT-4o	88.5	88.3	84.9	98.8	92.4	77.5
Doubao	86.2	81.7	84.9	91.6	86.5	82.1
GLM-4-Voice	82.3	77.7	88.9	93.7	82.7	80.0
VITA-Audio	82.5	77.9	82.8	91.8	89.8	81.0
MiniCPM-o	79.8	71.5	78.8	88.2	80.7	72.0
Step-Audio	81.8	79.4	77.8	86.8	77.6	80.6
Kimi-Audio	79.2	74.5	75.0	82.6	76.7	74.2
Qwen-Omni	82.1	74.3	75.8	92.2	69.9	69.0
Human	67.4	68.2	66.7	81.3	75.3	69.6
Moshi	54.5	46.5	45.5	70.4	65.8	64.1
AnyGPT	67.6	55.4	64.7	58.4	30.9	29.3

Table 4: Turn-level trends in response quality (↑) and content density (↑). T1–T3 denote 1–3 dialogue turns. Darker shades indicate degraded performance. See Appendix D.2 for calculation methods.[2]