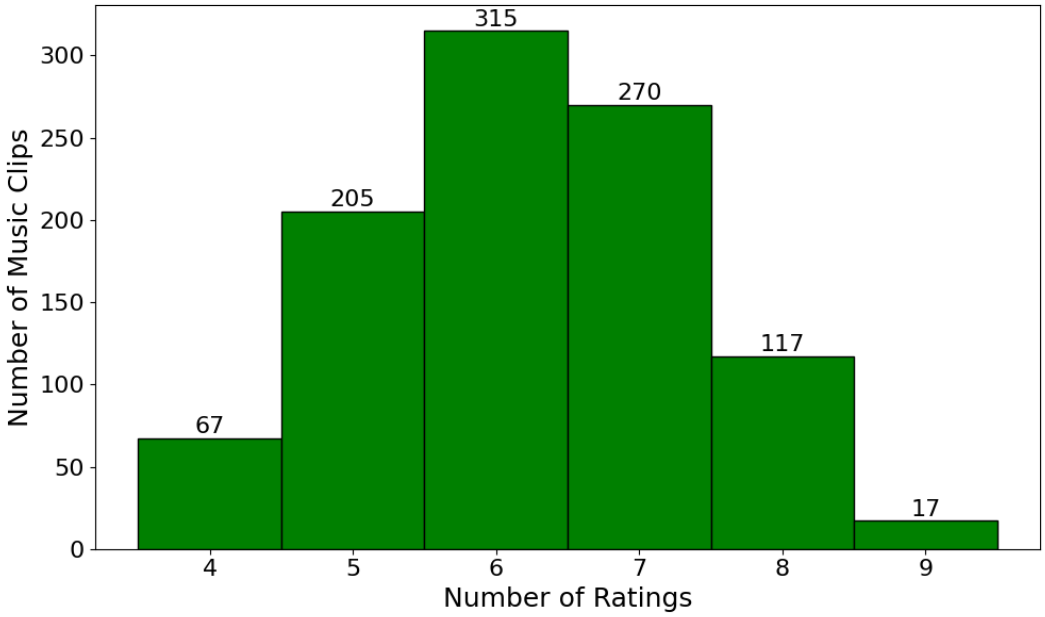

Figure S1. Distribution of the number of ratings per clip in AImoclips.

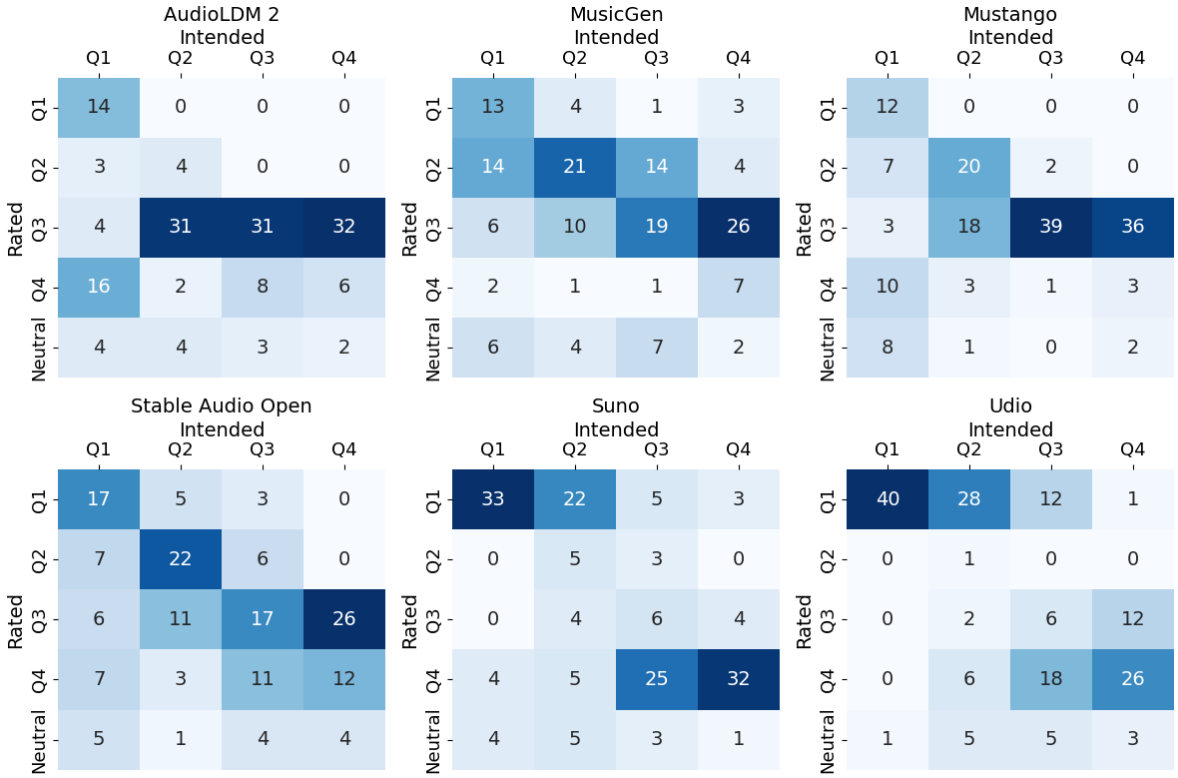

Figure S2. Confusion matrices by TTM system for quadrant-level emotion classification in the valence–arousal (VA) space (Q1: high valence–high arousal; Q2: low valence–high arousal; Q3: low valence–low arousal; Q4: high valence–low arousal). Clips with a mean valence or arousal rating exactly equal to 5.0 (the decision boundary between the quadrants) were categorized as "Neutral."

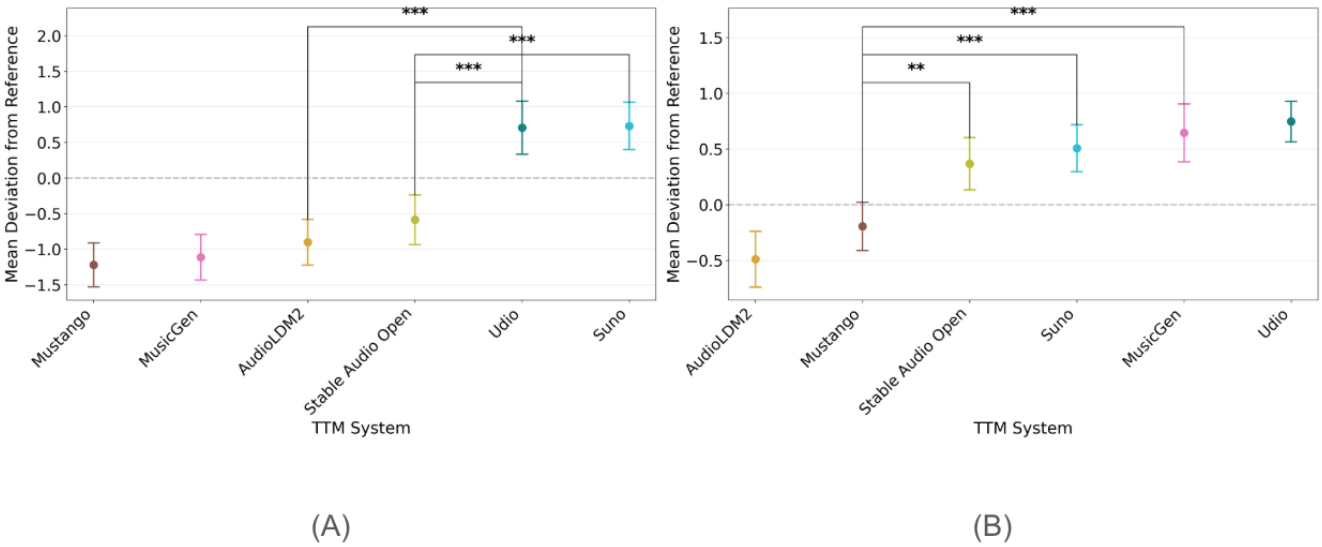

Figure S3. Mean deviation plots for each TTM system with 95% confidence intervals. Among all pairs with statistically significant differences, the three smallest differences are indicated. (*$p < .05$, **$p < .01$, ***$p < .001$.) (A) Mean valence deviations. (B) Mean arousal deviations.

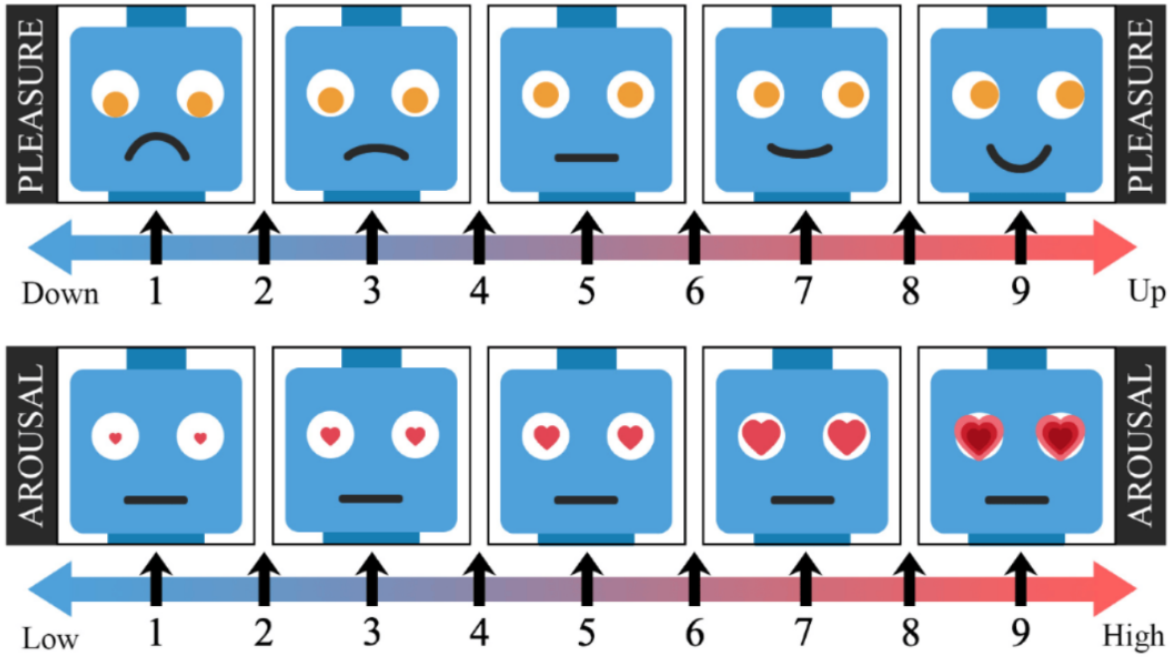

Figure S4. Self-Assessment Manikin (SAM) scale images[1] shown alongside the valence (top) and arousal (bottom) rating questions in the online survey.

Table S1. Overall and quadrant-wise summary statistics for valence and arousal ratings of clips in AImoclips.

Quadrant	Valence		Arousal		n
	Mean	Std	Mean	Std	
Overall	4.854	1.351	4.618	1.472	991
Q1	5.674	1.086	5.692	1.085	246
Q2	4.388	1.468	5.325	1.405	248
Q3	4.591	1.288	4.028	1.163	250
Q4	4.771	1.166	3.436	0.882	247

Table S2. Summary statistics of Euclidean distances between rated scores and reference scores, by quadrant.

Quadrant	Mean	Std	n
Q1	2.654	1.092	246
Q2	2.326	1.072	248
Q3	2.603	1.174	250
Q4	2.836	1.017	247

Table S3. Summary statistics of valence and arousal deviations from reference scores, by TTM system.

TTM System	Valence		Arousal		n
	Mean Deviation	Standard Error	Mean Deviation	Standard Error	
AudioLDM 2	-0.903	0.162	-0.490	0.127	164
MusicGen	-1.115	0.163	0.644	0.131	165
Mustango	-1.223	0.157	-0.196	0.110	165
Stable Audio Open	-0.588	0.177	0.366	0.119	167
Suno	0.730	0.168	0.505	0.107	164
Udio	0.707	0.189	0.746	0.092	166

Table S4. Summary statistics of Euclidean distances between rated scores (or reference scores) and the center of the valence–arousal plane ($v = 5$, $a = 5$), by TTM system. t and p denote the test statistic and p-value from two-sample t-tests comparing each system's clip ratings with the reference score set across all emotion intents.

TTM System	Mean	Std	t	p
Overall	1.857	0.840	-37.178	<.001
AudioLDM 2	1.843	0.792	-14.559	<.001
MusicGen	1.767	0.808	-15.509	<.001
Mustango	1.882	0.958	-13.490	<.001
Stable Audio Open	1.560	0.771	-19.449	<.001
Suno	2.118	0.792	-13.443	<.001
Udio	1.979	0.809	-16.408	<.001
Reference Scores	3.051	0.599	-	-

Table S5. Summary statistics for valence and arousal ratings, by TTM system and quadrant.

TTM System	Quadrant	Valence		Arousal		*n*
		Mean	Std	Mean	Std	
AudioLDM 2	Overall	4.339	1.160	3.875	0.992	41
	Q1	5.496	0.801	4.805	0.804	41
	Q2	3.391	1.021	3.962	1.046	41
	Q3	4.363	0.813	3.497	0.603	42
	Q4	4.099	0.888	3.227	0.668	40
MusicGen	Overall	4.157	1.167	4.971	1.309	41
	Q1	4.856	1.238	5.625	1.035	41
	Q2	3.757	1.235	5.635	1.145	40
	Q3	3.579	0.942	4.665	1.243	42
	Q4	4.432	0.744	4.004	1.052	42
Mustango	Overall	3.997	1.073	4.156	1.264	41
	Q1	5.161	0.577	5.217	0.701	40
	Q2	3.493	1.019	4.878	1.214	42
	Q3	3.500	0.826	3.284	0.874	42
	Q4	3.885	0.859	3.273	0.712	41
Stable Audio Open	Overall	4.690	1.087	4.733	1.302	41
	Q1	5.329	0.987	5.385	0.877	42
	Q2	4.166	1.085	5.605	1.332	42
	Q3	4.674	1.102	4.576	0.743	41
	Q4	4.588	0.858	3.363	0.785	42
Suno	Overall	5.967	1.015	4.867	1.776	41
	Q1	6.676	0.676	6.469	1.070	41
	Q2	5.504	1.159	5.873	1.437	41
	Q3	5.648	0.937	3.757	1.242	42
	Q4	6.047	0.817	3.357	0.913	40
Udio	Overall	5.974	0.957	5.100	1.648	41
	Q1	6.519	0.633	6.646	0.626	41
	Q2	5.991	0.929	5.993	1.150	42
	Q3	5.812	0.920	4.409	1.307	41
	Q4	5.583	1.061	3.372	0.906	42

References

1. He, X. & Song, N. Emotional value in online education: A framework for service touchpoint assessment. *Sustainability* **15**, DOI: 10.3390/su15064772 (2023).