

A APPENDIX

A.1 USE OF LARGE LANGUAGE MODELS

Large Language Models (LLMs) utilized in this work are as follows: (1) *Topics Initiation* during data processing of the Real-Pod dataset, which is elaborated in Section 3 (2) *LLM-as-a-Judge* method in text-based evaluation, which is illustrated in Section 4 (3) *Summarized Users' Justifications* in the Questionnaire-based MOS Test, which is described in Section 7.3 (Questionnaire-based MOS Test).

A.2 TEXT-BASED EVALUATION

A.2.1 QUANTITATIVE METRICS

Table 4: **GPT-4**: Quantitative metrics in text-based evaluation.

Metrics	Overall	Fiction	Education	Business	True Crime	Health & Fitness
Distinct_2	0.9619	0.9643	0.9588	0.9567	0.9689	0.9638
Info-Dens	6.4507	6.5865	6.4569	6.3213	6.6541	6.3880
Sem-Div	0.1293	0.1204	0.1115	0.1214	0.1443	0.1106
MATTR	0.6914	0.7027	0.6989	0.6933	0.6831	0.6870
Metrics	Sports	Comedy	History	News	TV & Film	Society & Culture
Distinct_2	0.9536	0.9633	0.9471	0.9486	0.9678	0.9659
Info-Dens	6.4228	6.2256	6.3792	6.3225	6.7614	6.6473
Sem-Div	0.1248	0.1356	0.1451	0.1208	0.1553	0.1507
MATTR	0.6973	0.6922	0.6905	0.6756	0.6903	0.6901
Metrics	Arts	Leisure	Music	Kids	Mental Health	Science & Tech
Distinct_2	0.9675	0.9729	0.9555	0.9559	0.9699	0.9710
Info-Dens	6.5054	6.5233	6.4119	6.2310	6.4787	6.3454
Sem-Div	0.1374	0.1117	0.1320	0.1229	0.1247	0.1286
MATTR	0.6885	0.7136	0.6677	0.6884	0.6994	0.6960

Table 5: **PodAgent**: Quantitative metrics in text-based evaluation.

Metrics	Overall	Fiction	Education	Business	True Crime	Health & Fitness
Distinct_2	0.9741	0.9743	0.9730	0.9758	0.9796	0.9825
Info-Dens	7.1767	7.3791	7.2163	7.1126	7.1810	7.2927
Sem-Div	0.1372	0.1384	0.1210	0.1254	0.1514	0.1171
MATTR	0.7216	0.7399	0.7291	0.7258	0.7263	0.7386
Metrics	Sports	Comedy	History	News	TV & Film	Society & Culture
Distinct_2	0.9678	0.9808	0.9483	0.9735	0.9782	0.9744
Info-Dens	7.1239	7.1600	7.1004	7.1282	7.3311	6.9568
Sem-Div	0.1487	0.1236	0.1543	0.1379	0.1690	0.1344
MATTR	0.7183	0.7248	0.6752	0.7156	0.7274	0.7119
Metrics	Arts	Leisure	Music	Kids	Mental Health	Science & Tech
Distinct_2	0.9701	0.9790	0.9739	0.9747	0.9815	0.9725
Info-Dens	7.1977	7.3227	7.0558	7.0930	7.1822	7.1711
Sem-Div	0.1283	0.1445	0.1440	0.1353	0.1259	0.1331
MATTR	0.7101	0.7275	0.7114	0.7279	0.7328	0.7249

Table 6: **MoonCast**: Quantitative metrics in text-based evaluation.

Metrics	Overall	Fiction	Education	Business	True Crime	Health & Fitness
Distinct_2	0.9128	0.9219	0.8998	0.8952	0.9132	0.9478
Info-Dens	6.0935	6.3613	5.9779	5.9931	6.0388	6.4230
Sem-Div	0.1326	0.1515	0.1079	0.1326	0.1405	0.1324
MATTR	0.6323	0.6598	0.6237	0.6310	0.6391	0.6698
Metrics	Sports	Comedy	History	News	TV & Film	Society & Culture
Distinct_2	0.9159	0.9232	0.9169	0.9047	0.9408	0.8959
Info-Dens	6.1933	6.1729	6.2229	5.9672	6.2855	5.8031
Sem-Div	0.1451	0.1311	0.1460	0.1176	0.1318	0.1282
MATTR	0.6402	0.6435	0.6276	0.6111	0.6595	0.6121
Metrics	Arts	Leisure	Music	Kids	Mental Health	Science & Tech
Distinct_2	0.9252	0.8889	0.8957	0.9039	0.9222	0.9073
Info-Dens	6.2335	5.9713	5.9411	5.9291	6.0298	6.0459
Sem-Div	0.1444	0.1309	0.1227	0.1291	0.1187	0.1432
MATTR	0.6370	0.6124	0.6035	0.6183	0.6321	0.6277

Table 7: **Real-Pod**: Quantitative metrics in text-based evaluation.

Metrics	Overall	Fiction	Education	Business	True Crime	Health & Fitness
Distinct_2	0.9200	0.9292	0.9275	0.9049	0.9169	0.9273
Info-Dens	8.1168	8.2849	8.1160	7.7755	8.5675	7.9301
Sem-Div	0.1677	0.1776	0.1579	0.1433	0.1906	0.1646
MATTR	0.6251	0.6313	0.6346	0.6041	0.6261	0.6380
Metrics	Sports	Comedy	History	News	TV & Film	Society & Culture
Distinct_2	0.9244	0.8994	0.9272	0.9100	0.9201	0.8932
Info-Dens	8.0993	8.2755	8.8282	7.7886	8.4005	7.7375
Sem-Div	0.1919	0.1660	0.1845	0.1618	0.1784	0.1701
MATTR	0.6434	0.5999	0.6304	0.6102	0.6363	0.5823
Metrics	Arts	Leisure	Music	Kids	Mental Health	Science & Tech
Distinct_2	0.9111	0.9242	0.9420	0.9092	0.9298	0.9439
Info-Dens	8.1093	7.6949	8.0925	7.7708	8.2119	8.3031
Sem-Div	0.1653	0.1591	0.1761	0.1492	0.1668	0.1485
MATTR	0.6063	0.6176	0.6513	0.6200	0.6373	0.6582

A.2.2 LLM-AS-A-JUDGE

Table 8: **LLM-as-a-Judge: comparison between GPT-4 and PodAgent**. Scores range from -3 to 3. Positive values indicate that PodAgent outperforms GPT-4; Negative values suggest the opposite.

Metrics	Overall	Fiction	Education	Business	True Crime	Health & Fitness
Coherence	0.7059	0.5000	0.8333	1.0000	1.0000	0.6667
Engagingness	1.0294	1.1667	1.0000	1.1667	0.6667	1.1667
Diversity	1.1765	1.3333	1.0000	1.3333	0.8333	1.5000
Informativeness	1.6078	1.5000	1.6667	2.0000	1.1667	1.6667
Speaker Difference	1.0637	0.9167	1.0000	1.1667	0.6667	1.0000
Overall	1.3064	1.2500	1.3333	1.6667	0.8333	1.2500
Metrics	Sports	Comedy	History	News	TV & Film	Society & Culture
Coherence	0.5000	0.8333	1.1667	0.6667	0.8333	0.1667
Engagingness	1.1667	1.5000	1.5000	0.6667	0.1667	0.6667
Diversity	1.1667	1.8333	1.5000	1.3333	1.3333	0.8333
Informativeness	1.5000	2.1667	1.5000	2.0000	1.3333	0.8333
Speaker Difference	1.1667	1.5000	1.1667	1.3333	1.1667	1.3333
Overall	1.5000	1.8333	1.5000	1.5000	0.8333	0.5000
Metrics	Arts	Leisure	Music	Kids	Mental Health	Science & Tech
Coherence	0.6667	0.5000	0.6667	0.5000	0.3333	1.1667
Engagingness	1.1667	1.1667	1.1667	1.0000	0.8333	1.3333
Diversity	1.1667	1.1667	1.0000	1.0000	0.3333	1.3333
Informativeness	1.8333	2.0000	1.8333	1.3333	1.1667	1.8333
Speaker Difference	1.3333	1.1667	0.8333	0.8333	0.6667	0.8333
Overall	1.5000	1.6667	1.5000	1.1667	0.8333	1.5417

A.3 SPEECH-BASED EVALUATION (SUBJECTIVE)

Welcome to the **podcast dialogue naturalness evaluation!**

Task Description:

In this test, you will listen to podcast dialogue segments generated by different systems. Your task is to evaluate the **naturalness** of the dialogue in each segment on a scale from **0 to 100**.

Evaluation Criteria:

- **0 – 20 (Bad):** The dialogue is completely unnatural, robotic, or awkward. It does not resemble a real conversation.
- **20 - 40 (Poor):** The dialogue has significant unnaturalness, with multiple awkward phrases, robotic tones, or inconsistent flows.
- **40 - 60 (Fair):** The dialogue is somewhat natural but has noticeable issues. It may feel rehearsed or lack smooth transitions.
- **60 - 80 (Good):** The dialogue is mostly natural, with minor unnatural elements. It resembles a real conversation but could still be improved.
- **80 - 100 (Excellent):** The dialogue sounds completely natural, like a real, spontaneous conversation between people.

Important Notes:

- The content of the dialogues may differ across systems. Please focus on the **overall naturalness** of the dialogue rather than the specific content or details (e.g., timbre, accent, noise or cut-off effects)
- In other words, **how realistic and similar are these dialogue segments to real podcast conversations?**
- Each test group includes a **Reference** audio extracted from a real podcast episode, representing the **"Excellent"** level of naturalness. This reference is provided to help calibrate your scoring.
- You can replay the audio segments as many times as you wish before assigning a score.
- Use headphones in a quiet environment for the best experience.

Your feedback is valuable. Thank you for participating!

Figure 7: Dialogue Naturalness Evaluation - Instruction page.

Instructions:

1. Listen to all the audio segments provided on this page.
2. Drag the slider below each audio to assign a score based on how natural the dialogue sounds.
3. After scoring all segments, click "Next" to proceed to the next page.

Evaluation Criteria:

- **0 – 20 (Bad):** The dialogue is completely unnatural, robotic, or awkward. It does not resemble a real conversation.
- **20 - 40 (Poor):** The dialogue has significant unnaturalness, with multiple awkward phrases, robotic tones, or inconsistent flows.
- **40 - 60 (Fair):** The dialogue is somewhat natural but has noticeable issues. It may feel rehearsed or lack smooth transitions.
- **60 - 80 (Good):** The dialogue is mostly natural, with minor unnatural elements. It resembles a real conversation but could still be improved.
- **80 - 100 (Excellent):** The dialogue sounds completely natural, like a real, spontaneous conversation between people.

Figure 8: Dialogue Naturalness Evaluation - Test page.

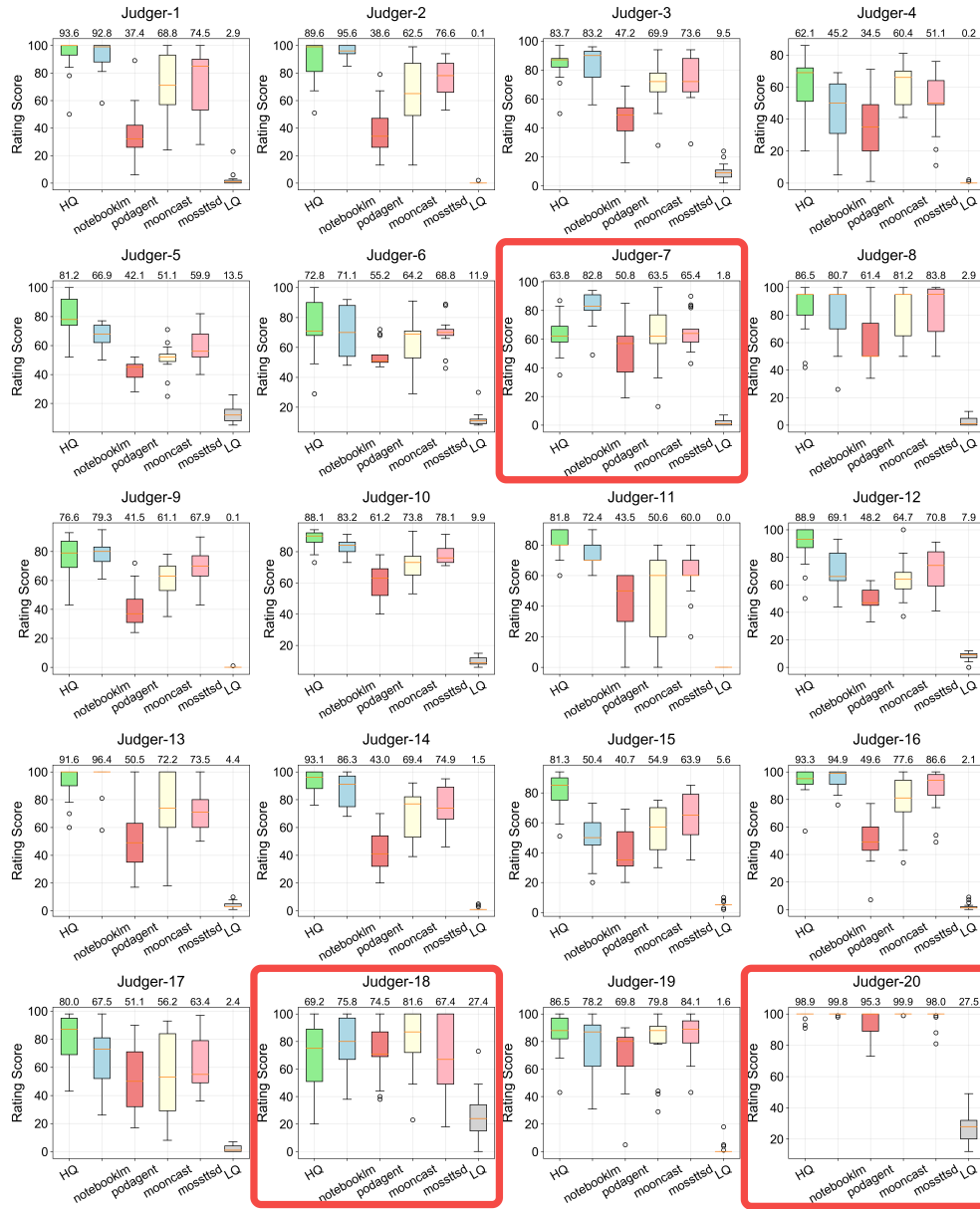


Figure 9: Dialogue Naturalness Evaluation test results from each judgers.

A.4 AUDIO-BASED EVALUATION (OBJECTIVE)

IDL: LOUD-IT; TP: LOUD-TP; LRA: LOUD-RA.

$$S_{\text{IDL}} = \begin{cases} 1, & -18 \leq \text{IDL} \leq -14, \\ e^{-k_1 \cdot (-18 - \text{IDL})}, & \text{IDL} < -18, \\ e^{-k_2 \cdot (\text{IDL} + 14)}, & -14 < \text{IDL}, \end{cases} \quad (2)$$

where k_1 is set as 0.0858 to ensure S_{IDL} is around 0.6 when $\text{IDL} = -23$, and k_2 is set as 0.3291 to make S_{IDL} close to 0 when $\text{IDL} = 0$.

$$S_{\text{TP}} = \begin{cases} 1, & \text{TP} \leq -1 \\ e^{-k_3 \cdot (\text{TP} + 1)}, & \text{TP} > -1 \end{cases} \quad (3)$$

where k_3 is set as 4.605 to ensure S_{TP} is close to 0 when TP approaches 0.

$$S_{\text{LRA}} = \begin{cases} 1, & 4 \leq \text{LRA} \leq 18, \\ e^{-k_4 \cdot (4 - \text{LRA})}, & \text{LRA} < 4, \\ e^{-k_5 \cdot (\text{LRA} - 18)}, & \text{LRA} > 18. \end{cases} \quad (4)$$

where k_4 is set as 1.1513 to ensure S_{LRA} approaches 0 when $\text{LRA} = 0$, and k_5 is set as 0.2554 to ensure $S_{\text{LRA}} \approx 0.6$ when $\text{LRA} = 20$.

Table 9: Audio-based objective metrics - Quantitative scores.

System	LOUD_IT_SCORE	LOUD_TP_SCORE	LOUD_LRA_SCORE	SMR_BASIC_SCORE	CASP
Real-Pod	0.72	0.53	0.82	0.99	0.58
PodAgent	0.80	0.32	1.00	1.00	0.56
MoonCast	1.00	0.01	0.68	-	-
Muyan-TTS	0.88	1.00	0.83	-	-
Dia	0.98	0.01	0.95	-	-
MOSS-TTSD	0.88	0.02	0.99	-	-
NotebookLM	0.51	0.56	1.00	-	-

A.5 AUDIO-BASED EVALUATION (SUBJECTIVE)

A.5.1 PILOT TEST

Section 1: Quantitative Analysis (0-10 Scale)

0 = not met at all, 5 = moderately met, 10 = fully met. Comments are optional but encouraged.

1. How well does the tone of the host or guest suit the podcast content?

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

2. How clearly and effectively do the speakers deliver the podcast content?

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

3. Is the speaking pace appropriate and easy to follow?

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

4. How engaging and enjoyable is the podcast? (Does it sustain your attention throughout the episode?)

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

5. How satisfied are you with the podcast's audio quality? (e.g., clarity, background noise)

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

6. If background music or sound effects are present, how well do they enhance rather than interfere with the content? (Select 5 if there is no background music or sound effects)

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

7. How likely are you to want to listen to the full episode after hearing this excerpt?

not met at all	0	1	2	3	4	5	6	7	8	9	10	fully met
	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	☆	

Section 2: Qualitative Analysis (YES/NO/CAN'T TELL)

8. Does the podcast include a clear introduction and conclusion?

YES	NO	CAN'T TELL
☆	☆	☆

9. Are background music or sound effects present in the podcast?

YES	NO	CAN'T TELL
☆	☆	☆

10. Does the podcast sound like it was created by humans rather than AI? ("Yes" = more like humans, "No" more like AI)

YES	NO	CAN'T TELL
☆	☆	☆

Figure 10: Questionnaire-based MOS test - Pilot test version.

A.5.2 QUESTIONNAIRE-BASED MOS TEST

Experiment Settings: Lengthy listening tests can be exhausting and may lead to inaccurate feedback. It is essential to ensure the overall test duration does not exceed 30 minutes. In the Questionnaire-based MOS Test, each audio sample is around 3 minutes and requires answering 10 questions with corresponding justifications. Based on the Dialogue Naturalness Test results shown in Figure 7.2, we selected 4 representative systems. Each test group included four podcast samples from different systems but within the same podcast category. According to actual test results, each group took an average of 24 minutes to complete. The 4 representative systems are:

- **PodAgent:** An open-source podcast generation framework incorporating conversation script generation, automatic voice selection, speech synthesis, and BMSE enhancement.
- **MOSS-TTSD:** Achieved the highest score among the open-source systems utilized in the Dialogue Naturalness Evaluation (Figure 7.2).
- **NotebookLM:** A pioneering podcast generation product, widely recognized for its exceptional performance, is nearly indistinguishable from real podcasts.
- **Real-Pod:** A collection of podcasts sourced from the real world.

Welcome to the Podcast Evaluation Questionnaire!

Study Description:

In this study, we aim to collect **authentic feedback** on podcast audio clips. You will listen to **4 different podcast audio files**, each discussing potentially different topics. The primary goal of this research is to evaluate the **overall production quality** of the podcast segments, rather than the specific content or themes being discussed.

Each audio clip is approximately **3 minutes long** and is constructed by combining three key segments from a full podcast episode:

<The **first** minute | The **middle** minute | The **final** minute>

A brief notification sound will indicate the transitions between these segments.

About the questionnaire:

It consists of 8 questions, which are designed to assess the podcast audio across multiple dimensions, such as:

- Speaker's expression / Information delivery
- Audio quality / engagingness / music or sound effect harmony

Notice:

- We kindly ask you to **avoid** rating based on the **discussion** topic and instead focus on the requested dimension.
- Please **listen to each audio carefully**, ideally using **headphones** for optimal clarity.
- **Incomplete or insincere responses** may be subject to return. We kindly ask you to provide thoughtful and genuine feedback to ensure the effectiveness of this study.
- Please enter your **Prolific ID** as the "**Username**" in the final submission page.

Your feedback is **extremely valuable**. Thank you for your participation!

Figure 11: Questionnaire-based MOS test - Final version - Instruction page.

1134

1135

1136

1137 Q1. How many speakers are there in the podcast?

1138

1139

1140 Q2. How satisfied are you with the podcast's audio quality (e.g., clarity, volume levels, background noise)?

1141 ☐ 1 = Very dissatisfied

1142 ☐ 2 = Dissatisfied

1143 ☐ 3 = Neutral

1144 ☐ 4 = Satisfied

1145 ☐ 5 = Very satisfied

1146

1147 Why? (Required but can be simple. The same requirement for other "Why?" questions.)

1148

1149

1150

1151 Q3. Do you like the way the guests and hosts express themselves?

1152 ☐ 1 = Strongly dislike it

1153 ☐ 2 = Dislike it

1154 ☐ 3 = Neutral

1155 ☐ 4 = Like it

1156 ☐ 5 = Love it

1157

1158 Why?

1159

1160

1161

1162 Q4. Do you think the speakers are effectively delivering the information?

1163 ☐ 1 = Not at all effectively

1164 ☐ 2 = Not very effectively

1165 ☐ 3 = Neutral

1166 ☐ 4 = Somewhat effectively

1167 ☐ 5 = Very effectively

1168

1169 Why?

1170

1171

1172

1173 Q5. If music or sound effects are present, do they enhance or interfere with the content? (Select Neutral if none are present)

1174 ☐ 1 = Greatly interfere

1175 ☐ 2 = Somewhat interfere

1176 ☐ 3 = Neutral

1177 ☐ 4 = Somewhat enhance

1178 ☐ 5 = Greatly enhance

1179

1180 Why?

1181

1182

1183

1184

1185

1186

1187

Figure 12: Questionnaire-based MOS test - Final version (Question 1-5).

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

Q6. How **engaging** is the podcast?

☐

1 = Not engaging at all

☐

2 = Slightly engaging

☐

3 = Neutral☐☐

Why?

Q7. How likely are you to listen to the **full episode** after hearing this?

☐

1 = Not likely at all☐☐☐☐

Why?

Q8. Does the podcast sound like it was created by **humans** rather than **AI**?

☐

1 = Definitely AI☐☐☐☐

Why?

Q9. (Optional) Any additional comments on this podcast audio?

Figure 13: Questionnaire-based MOS test - Final version (Question 6-9).

Table 10: Questionnaire-based MOS test - (Q.) represents the average score from the direct scoring answers, and (J.) represents the score derived from the justifications.

Metrics \ Systems	MOSS-TTSD		NotebookLM		PodAgent		Real-Pod	
	Q.	J.	Q.	J.	Q.	J.	Q.	J.
Information Delivery	4.0	3.0	4.2	4.0	1.6	1.0	4.2	4.0
Music/Sound Effects	N/A	N/A	N/A	N/A	2.4	2.0	3.3	3.0
Engagement Level	2.2	3.0	3.1	3.0	1.1	1.0	3.6	4.0
Full Episode Likelihood	2.1	2.0	2.1	3.0	1.0	1.0	2.3	3.0
Human Likelihood	3.0	3.0	3.3	3.5	1.1	1.0	4.2	4.0
Audio Quality	3.5	3.0	4.2	4.0	3.0	2.0	3.9	4.0
Speaker Expression	3.3	3.0	4.0	3.0	1.5	1.0	3.4	4.0

A.6 SYSTEM ANALYSIS REPORT



Figure 14: System analysis report based on PodEval - PodAgent.