**Human Assessment using Scoring Rubric**

Product: Apple AirPod Max

Rater: human3

Author: LLM

1.1 Coverage of overall contents and trends (10 pts, 2.5 pts per item)

Are each item(general sentiment, positive aspects and main complaints, and key insights) well-summarized?

1.2 Clarity and conciseness (10 pts)

Is the information well-synthesized, free of redundancy, and clearly written?

2. Sentiment Overview (15 pts)

2.1 Accuracy of sentiment statistics (15 pts, 3 pts per item)

Does the breakdown of positive, neutral, and negative reviews, average rating, and NPS score accurately reflect the data?

3. Insights by 3 Product Features (20 pts)

3.1 Coverage of key features (10pts)

Does it cover the 3 key product features mentioned in reviews with clear positive/negative aspects? (Base 1pts, 3 pts per item)

3.2 Depth of insight and examples (10pts)

Does it include trend-based analysis and proper review quotes with in-depth observations? (Base 1pts, 3 pts per item)

4. Pros and Cons Summary

4.1 Clear and specific points (7pts)

Does it include clear and detailed pros and cons that are supported by the reviews?

4.2 Coverage and representativeness of aspects (8 pts)

Do the listed pros and cons reflect the most important and frequently mentioned aspects from the review dataset?

5. Recommendations

5.1 Review-based suggestions (10 pts)

Do the recommendations clearly come from review insights and are well-grounded in patterns identified?

5.2 Clarity and actionability (10 pts)

Are suggestions specific, actionable, and related to the product and business?

6. Overall Structure and Style (10 pts)

6.1 Template adherence and section formatting (5 pts)

Does the report strictly follow the template layout, including section headings, order, and word count ranges?

6.2 Tone, grammar, and consistency (5 pts)

Is the language clear, professional, and grammatically correct with no stylistic inconsistencies?

7. Qualitative Reviewer Comment (Only for the qualitative review of the LLM-generated report)

In addition to the numerical scores, please write an overall comment about the LLM-generated report, including your opinions and impressions that may not be reflected enough in the rubric scores. You may highlight strengths, weaknesses, and what stood out to you. (Word count: 50-100 words)

|  |  |  |
| --- | --- | --- |
| 1. Executive Summary (20 pts) | 1.1 Coverage of overall contents and trends (10 pts, 2.5 pts per item) | 7 |
| 1.2 Clarity and conciseness (10 pts) | 6 |
| 2. Sentiment Overview (15 pts) | 2.1 Positive reviews | 0 |
| 2.2 Neutral reviews | 0 |
| 2.3 Negative reviews | 0 |
| 2.4 Average rating | 3 |
| 2.5 Net Promoter Score | 3 |
| 3. Insights by 3 Product Features (20 pts) | 3.1 Coverage of key features (10pts) | 7 |
| 3.2 Depth of insight and examples (10pts) | 9 |
| 4. Pros and Cons Summary | 4.1 Clear and specific points (7pts) | 7 |
| 4.2 Coverage and representativeness of aspects (8 pts) | 8 |
| 5. Recommendations | 5.1 Review-based suggestions (10 pts) | 8 |
| 5.2 Clarity and actionability (10 pts) | 5 |
| 6. Overall Structure and Style (10 pts) | 6.1 Template adherence and section formatting (5 pts) | 5 |
| 6.2 Tone, grammar, and consistency (5 pts) | 5 |

Total: 73

7. Qualitative Reviewer Comment (Only for the qualitative review of the LLM-generated report)

In addition to the numerical scores, please write an overall comment about the LLM-generated report, including your opinions and impressions that may not be reflected enough in the rubric scores. You may highlight strengths, weaknesses, and what stood out to you:

The LLM was good at extracting lots of information in a short time, but it missed some information too. For example, design was mentioned several times but LLM didn't put it in Section 3. And it made different decisions in Sentiment analysis. Recommendations indicate that like it doesn't have any idea about real business. But it was really good at following the format and grammar.