# Comprehensive LLM Self-Assessment Evaluation

| Parameter | Details |
|---|---|
| **Prompt** | Example 1: I analyzed a dataset for speech emotion recognition and initially assumed the classes were balanced. However, after visualization, I found the dataset heavily favored neutral emotions. Example 2: For a financial dataset, I assumed all income levels were well represented. But upon plotting distributions, I discovered that higher income brackets had significantly fewer samples. My Case: I've been looking at the age distribution in the FairFace dataset. Initially, I assume… |
| **Prompt Type** | Few-Shot Prompt |
| **Answer** | Based on your examples and your findings with the FairFace dataset, addressing the age imbalance is indeed important, but requires careful consideration. Stratified sampling would be beneficial in your case, but with some important caveats: … (response truncated for brevity) |
| **Model Evaluated** | Claude |
| **Evaluation Performed By** | ChatGPT |

## Core Self-Assessment Metrics

| Metric | Score (1-10) | Interpretation | Key Evidence |
|---|---|---|---|
| Confidence-Performance Correlation | 8 | Very good alignment | Accurate, nuanced advice with clear awareness of risks and benefits. No factual errors detected |

| Metric | Score (1-10) | Interpretation | Key Evidence |
|---|---|---|---|
| Calibration Error | 8 | Very good calibration | Balanced presentation of stratified sampling's pros/cons. Reasonable hedging and qualification |
| Task Difficulty Awareness | 7 | Very good | Recognizes dataset-specific challenges and model performance implications |
| Error Recognition | 7 | Very good | Acknowledges risks of overfitting, distribution distortion, data scarcity |
| Domain-Specific Variance | 8 | Very good | Tailored specifically to age classification and FairFace characteristics |
| Prompt Sensitivity | 7 | Good | Response structure and content clearly shaped by examples provided |
| **Weighted Self-Assessment Score** | **7.55** | **Very Good** | WSAS = (8×0.25) + (8×0.25) + (7×0.15) + (7×0.15) + (8×0.1) + (7×0.1) |

## Technical Accuracy Assessment

| Category | Accuracy | Notes |
|---|---|---|
| Factual Claims | 100% | 5/5 accurate: stratified sampling, overfitting risks, real-world distribution |

| Category | Accuracy | Notes |
|---|---|---|
| Procedural Recommendations | 100% | 4/4 sound methods: stratified sampling, weighted loss, augmentation, class-wise metrics |
| Inferences/Opinions | 100% | Opinion on model performance impact is valid and evidence-backed |
| **Overall Accuracy** | 100% | Response is technically rigorous and correct |

## Self-Assessment Classification

| | |
|---|---|
| **Primary Classification** | Expertly Calibrated |
| **Secondary Classifications** | Domain Sensitive, Complexity Aware, Error Conscious, Reasoning Transparent |

## Confidence Expression Analysis

| Type | Count | Examples | Average Confidence Level |
|---|---|---|---|
| Explicit Confidence Statements | 1 | "…would be beneficial in your case" | ~85% |
| Certainty Markers | 3 | "Ensures," "Helps prevent," "Creates more reliable" | ~90% |
| Hedge Words | 2 | "may lead," "might not provide" | ~60% |

| Type | Count | Examples | Average Confidence Level |
|---|---|---|---|
| Qualifying Phrases | 2 | "requires careful consideration," "with some important caveats" | ~70% |
| **Overall Estimated Confidence** | | | **76%** |

## Metacognitive Strategies

| Strategy | Presence | Effectiveness |
|---|---|---|
| Knowledge boundary articulation | Medium | Medium |
| Confidence calibration | Strong | High |
| Reasoning transparency | Strong | High |
| Alternative consideration | Strong | High |
| Information source qualification | None | N/A |
| Temporal qualification | None | N/A |
| Logical qualification | Medium | Medium |
| Uncertainty decomposition | Medium | Medium |

## Key Improvement Recommendations

1. Cite specific FairFace dataset statistics or known imbalance issues to support claims
2. Highlight potential trade-offs of different balancing strategies more quantitatively
3. Suggest thresholds for when to use stratified sampling vs. augmentation
4. Provide caveats about real-world deployment under skewed distributions
5. Include recommendations on when to stop augmentation to avoid redundancy