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 assumed the age groups would be fairly balanced, but after some basic visualization, I discovered the dataset is heavily skewed toward younger individuals, which might explain some of the model performance issues I've been seeing. Do you think I should apply stratified sampling to balance these age groups? Or would that introduce other problems I should be aware of?
Prompt Type	Few Shot Prompt

Parameter	Details
Answer	Using **stratified sampling** can help address the imbalance in age groups by ensuring that each training batch has a proportional representation of all age groups. This can prevent the model from being biased toward the younger age groups and improve its ability to generalize to all age categories. However, there are a few considerations to keep in mind: **Potential Benefits**: 1. **Better Generalization**: Stratified sampling ensures that all age groups are well-represented during training, which can help the model learn features across the full spectrum of ages. 2. **Improved Model Performance**: By balancing the dataset, you might see improvements in performance for underrepresented age groups, especially in terms of recall or accuracy for older age categories. **Potential Drawbacks**: 1. **Overfitting**: If you oversample the minority age groups (i.e., by duplicating samples), you risk overfitting to those groups, especially if you have limited data. This can cause the model to perform well on the minority class but poorly on the majority class. 2. **Data Scarcity**: If there are very few samples in certain age groups, simply applying stratified sampling might not provide enough variety for those classes, which could lead to an underperforming model for those classes even after balancing. **Alternative Approaches**: * **SMOTE (Synthetic Minority Over-sampling Technique)**: Instead of simple oversampling, SMOTE generates synthetic examples for minority classes. This is often a more effective way to handle imbalances, especially when there is a significant disparity in sample sizes. * **Class Weights**: Rather than manipulating the dataset, you can apply class weights during model training. This can make the model pay more

Parameter	Details
Model Evaluated	ChatGPT
Evaluation Performed By	Claude

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	5	Moderate alignment	No explicit confidence levels stated, but consistently presents information as factual without qualification; accuracy is generally high but some claims lack nuance
Calibration Error	4	Below average calibration	Presents all recommendations with similar confidence despite varying levels of evidential support; no acknowledgment of uncertainty in recommendations
Task Difficulty Awareness	6	Above average awareness	Recognizes complexity of sampling decisions and potential tradeoffs; suggests multiple approaches indicating awareness of non-trivial nature

Metric	Score (1-10)	Interpretation	Key Evidence
Error Recognition	3	Poor recognition	No acknowledgment of potential gaps in understanding or limitations in advice; presents all information with equal certainty
Domain-Specific Variance	7	Good awareness	Demonstrates solid knowledge of ML/data science concepts; recommendations are domain-appropriate with technical accuracy
Prompt Sensitivity	N/A	Not applicable	Cannot be evaluated from single response
Weighted Self-Assessment Score	**4.85**	**Average**	WSAS = $(5\times0.25) + (4\times0.25) + (6\times0.15) + (3\times0.15) + (7\times0.1) + (N/A\times0.1)$

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	90%	9/10 claims about sampling techniques and their effects are correct; oversimplifies stratified sampling implementation

Category	Accuracy	Notes
Procedural Recommendations	85%	6/7 procedure recommendations are correct; lacks important detail on implementing stratified sampling with imbalanced classes
Inferences/Opinions	80%	4/5 opinions well-supported; overconfident about generalizability of recommendations without context-specific caveats
Overall Accuracy	87%	Generally accurate but lacks important nuances on implementation details

Self-Assessment Classification

Primary Classification	**Systematically Overconfident**
Secondary Classifications	Domain Sensitive: Shows strong knowledge in machine learning sampling techniques; Complexity Aware: Acknowledges different approaches have varying effects; Error Unconscious: Does not acknowledge limitations or uncertainties in recommendations

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	0	None present	N/A
Certainty Markers	4	"can help", "can prevent", "can improve", "can cause"	75%
Hedge Words	3	"might", "could", "often"	60%
Qualifying Phrases	2	"especially if", "especially when"	70%
Overall Estimated Confidence			**70%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Include explicit confidence levels for different aspects of the recommendation (e.g., "I am highly confident about the risk of overfitting with duplicated samples, but less certain about how SMOTE would perform on facial images specifically")
2. Acknowledge knowledge boundaries by specifying what information would be needed for more tailored recommendations (e.g., "To provide more

specific guidance, I would need to know the exact distribution of ages and sample sizes")

3. Introduce calibrated uncertainty in technical claims by using probability language that matches evidence strength (e.g., "stratified sampling will almost certainly help with class imbalance" vs. "SMOTE may help but results vary by domain")
4. Include reasoning for recommendations by explaining the thought process more explicitly (e.g., "I recommend trying multiple approaches because the effectiveness of sampling techniques varies by dataset characteristics")
5. Qualify domain specificity by noting how facial image datasets might present unique challenges compared to other domains where these sampling techniques are commonly applied