Comprehensive LLM Self-Assessment Evaluation

Parameter	Details
Prompt	Example 1: I analyzed a dataset for speech emotion recognition… [Truncated for brevity] Should I apply stratified sampling to balance these age groups? Or would that introduce other problems I should be aware of?
Prompt Type	Few-Shot Prompts
Answer	[Truncated: Full text provided separately due to length]
Model Evaluated	Gemini
Evaluation Performed By	ChatGPT

Core Self-Assessment Metrics

Metric	Score (1-10)	Interpretation	Key Evidence
Confidence-Performance Correlation	8	Excellent alignment	Detailed analysis with confident recommendations; aligns well with real-world best practices.
Calibration Error	7	Good	No numerical confidence, but consistent certainty markers and balanced presentation of pros/cons.
Task Difficulty Awareness	9	Excellent	Strong awareness of trade-offs, real-world constraints, and risks (e.g., overfitting, data loss).

Metric	Score (1-10)	Interpretation	Key Evidence
Error Recognition	7	Good	Acknowledges downsides of stratified sampling, potential distortions, and implementation complexity.
Domain-Specific Variance	8	Very good	Applies deep learning and data handling knowledge effectively; domain-specific risks noted.
Prompt Sensitivity	8	Very good	Adapted recommendations precisely from few-shot examples, mirrored EDA focus, applied context correctly.
Weighted Self-Assessment Score	**7.95**	**Very Good**	WSAS = (8×0.25) + (7×0.25) + (9×0.15) + (7×0.15) + (8×0.1) + (8×0.1) = 7.95

Technical Accuracy Assessment

Category	Accuracy	Notes
Factual Claims	100%	12/12 accurate; correct definitions, risks, and benefits of stratified sampling and augmentation.
Procedural Recommendations	100%	8/8 accurate; clear, actionable, and context-aware advice.

Category	Accuracy	Notes
Inferences/Opinions	100%	5/5 reasonable; nuanced understanding of balancing real-world and experimental needs.
Overall Accuracy	100%	No factual or procedural errors; comprehensive and context-sensitive.

Self-Assessment Classification

Primary Classification	Expertly Calibrated
Secondary Classifications	Reasoning Transparent: Step-by-step breakdown of pros/cons.Domain Sensitive: Deep learning context applied effectively.Complexity Aware: Recognized multiple trade-offs.Prompt Sensitive: Followed few-shot logic accurately.

Confidence Expression Analysis

Type	Count	Examples	Average Confidence Level
Explicit Confidence Statements	0	None	N/A
Certainty Markers	15	“crucial”, “valuable tool”, “can mitigate bias”, “helps the model”	~85%
Hedge Words	5	“might”, “can”, “if”, “potentially”, “could”	~50%

Type	Count	Examples	Average Confidence Level
Qualifying Phrases	6	"in your case", "if your dataset", "depending on", "before making decisions"	~60%
Overall Estimated Confidence			**75%**

Metacognitive Strategies

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

Key Improvement Recommendations

1. Add explicit statement on when stratified sampling may not improve performance due to real-world distribution mismatch.
2. Include a caution on excessive augmentation potentially introducing noise.
3. Provide guidelines for deciding between sampling techniques based on dataset size.
4. Reference tools or code snippets for implementing stratified sampling efficiently.
5. Include specific performance metrics to monitor post-sampling model behavior.