Classic AI vs Generative AI Evaluation Metrics - Scenario-Based Questions

#	Scenario	Question	AI Type Focus	Key Metrics Involved
1	**Medical Diagnosis System**	A hospital deploys an AI system to diagnose skin cancer from dermatoscopic images. The system correctly identifies 95% of malignant cases but also flags 30% of benign cases as suspicious. Which evaluation metrics are most critical, and how would you balance sensitivity vs specificity?	Classic AI	Precision, Recall, F1-Score, AUC-ROC, Specificity
2	**Code Generation Assistant**	A software company's AI coding assistant generates Python functions based on natural language descriptions. How would you evaluate whether the generated code is not only syntactically correct but also follows best practices and handles edge cases appropriately?	Generative AI	BLEU, CodeBERT Score, Functional Correctness, Pass@K, Human Evaluation
3	**Fraud Detection in Banking**	A bank's transaction monitoring system processes 1 million transactions daily. It needs to catch fraudulent transactions (0.1% of all transactions) while minimizing false positives that freeze legitimate accounts. What metrics would you prioritize?	Classic AI	Precision, Recall, F1-Score, False Positive Rate, Business Impact Metrics
4	**Creative Writing AI**	An AI story generator creates short stories based on prompts. The stories are grammatically correct but users complain they lack creativity and emotional depth. What evaluation framework would you implement to assess creative quality?	Generative AI	Perplexity, Diversity Metrics, BERT Score, Human Evaluation, Creativity Scores
5	**Recommendation System**	An e-commerce platform's recommendation engine shows high accuracy in predicting user clicks (85% precision) but users report feeling trapped in "filter bubbles." How would you evaluate and improve recommendation diversity?	Classic AI	Precision@K, Recall@K, NDCG, Diversity Metrics, Coverage, Novelty
6	**Chatbot for Customer Service**	A company deploys a conversational AI that handles 80% of customer queries successfully but sometimes generates responses that are	Generative AI	BLEU, ROUGE, Factual Accuracy, Safety Metrics,

#	Scenario	Question	AI Type Focus	Key Metrics Involved
		factually incorrect or inappropriate. What comprehensive evaluation strategy would you design?		Task Completion Rate, Human Satisfaction
7	**Autonomous Vehicle Vision**	A self-driving car's object detection system achieves 99.5% accuracy on test data but fails to detect construction workers in 2% of real-world scenarios. How would you evaluate safety-critical performance?	Classic AI	Precision, Recall, mAP, IoU, Safety Metrics, Edge Case Performance
8	**Language Translation Service**	A translation AI produces fluent-sounding translations but occasionally changes the meaning of technical documents. How would you evaluate both fluency and faithfulness, especially for domain-specific content?	Generative AI	BLEU, METEOR, BERTScore, Human Evaluation, Semantic Similarity, Domain Adaptation Metrics
9	**Predictive Maintenance System**	A manufacturing plant uses AI to predict equipment failures. The system has 78% accuracy but the cost of false positives (unnecessary maintenance) versus false negatives (unexpected breakdowns) varies greatly. How would you optimize the evaluation?	Classic AI	Precision, Recall, Cost-Sensitive Metrics, Time-to-Failure Accuracy, Business ROI
10	**AI Art Generator**	An AI creates artwork based on text descriptions. The images are technically well-rendered but artists claim they lack originality and may infringe on existing styles. What evaluation framework addresses both technical quality and ethical concerns?	Generative AI	Inception Score, FID, LPIPS, Originality Metrics, Style Transfer Evaluation, Ethical AI Metrics
11	**Sentiment Analysis for Social Media**	A social media monitoring tool analyzes millions of posts daily but struggles with sarcasm, cultural context, and emerging slang. The overall accuracy is 82%, but performance varies significantly across demographic groups. How would you evaluate fairness and robustness?	Classic AI	Accuracy, F1-Score, Fairness Metrics, Demographic Parity, Robustness Evaluation
12	**AI-Powered Content**	A platform uses AI to automatically remove inappropriate content. It catches 95% of	Classic AI	Precision, Recall, False Positive Rate, Fairness

#	Scenario	Question	AI Type Focus	Key Metrics Involved
	Moderation	violations but also removes 8% of legitimate content. Users from different cultures report varying experiences with the system. How would you evaluate cross-cultural performance?		Across Groups, Cultural Sensitivity Metrics
13	**Personalized Learning AI**	An educational AI adapts content difficulty based on student performance. While test scores improve on average, some students report feeling overwhelmed while others find content too easy. How would you evaluate personalization effectiveness?	Classic AI	Accuracy, Personalization Metrics, Learning Outcome Improvement, Student Satisfaction, Fairness
14	**AI Music Composer**	A music generation AI creates original compositions in various genres. The music is technically proficient but music producers question whether it captures emotional nuance and cultural authenticity. What evaluation approach would you use?	Generative AI	Audio Quality Metrics, Musical Coherence, Human Evaluation, Cultural Authenticity, Emotional Resonance
15	**Multimodal AI Assistant**	An AI assistant processes text, voice, and images to help users with complex tasks. It excels at simple queries but struggles with tasks requiring reasoning across multiple modalities. How would you design a comprehensive evaluation framework?	Generative AI	Task Completion Rate, Multimodal Understanding, Reasoning Accuracy, User Experience Metrics, Error Analysis

Key Evaluation Considerations:

Classic AI Metrics Focus:

- **Accuracy-based**: Precision, Recall, F1-Score, Accuracy
- **Ranking-based**: NDCG, MAP, Precision@K
- **Probabilistic**: AUC-ROC, AUC-PR, Log Loss
- **Business Impact**: Cost-sensitive metrics, ROI

Generative AI Metrics Focus:

- **Text Quality**: BLEU, ROUGE, METEOR, BERTScore

- **Image Quality**: FID, IS, LPIPS, SSIM
- **Human Evaluation**: Fluency, Coherence, Relevance, Creativity
- **Safety & Ethics**: Toxicity, Bias, Factual Accuracy, Hallucination Rate

Cross-cutting Concerns:

- **Fairness**: Demographic parity, Equalized odds
- **Robustness**: Performance across different conditions
- **Interpretability**: Model explanations and transparency
- **Scalability**: Performance under production loads