SCHEDA DI VALUTAZIONE COMPLETA DEI MODELLI FINE-TUNED

---

1. Valutazione Qualitativa

- Accuratezza percepita: giudizio umano sulla correttezza.
- Stile e tono: coerenza linguistica.
- Pertinenza: risposta centrata sulla domanda.
- Regressioni (catastrophic forgetting): verifica su competenze generali.
- Robustezza: resistenza ad input rumorosi o ambigui.

---

2. Metriche Quantitative – Classificazione

- Accuracy: percentuale totale di predizioni corrette.
- Precision: riduce i falsi positivi.
- Recall (Sensitivity): riduce i falsi negativi.
- Specificity: capacità di riconoscere correttamente i negativi.
- F1-score: bilanciamento tra precision e recall.
- Confusion Matrix: distribuzione degli errori.
- ROC Curve & AUC-ROC: qualità del modello a varie soglie.
- AUC-PR: preferibile in dataset sbilanciati.
- Balanced Accuracy: media tra sensitivity e specificity.
- MCC: metrica robusta per classificazione.
- Cohen’s Kappa: accordo corretto per il caso.
- Log-Loss / Cross-entropy: qualità delle probabilità.
- Brier Score: calibrazione delle probabilità.

---

3. Metriche per Generazione Testo

- ROUGE: sovrapposizione lessicale.
- BLEU: n-gram corrispondenti.
- METEOR: sinonimi e stemming.
- BERTScore: similarità semantica.
- MoverScore: distanza semantica.
- Perplexity: sorpresa del modello.

---

4. Question Answering

- Exact Match (EM): corrispondenza testuale perfetta.
- F1 token-level: sovrapposizione parziale.

---

5. Ragionamento e Allucinazioni

- Hallucination Rate: quantità di errori fattuali.
- Truthfulness: aderenza ai fatti.
- Chain-of-Thought Faithfulness: coerenza del ragionamento.
- Pass@k: qualità nella generazione codice.

---

6. Sicurezza e Bias

- Toxicity Score: rischio linguaggio offensivo.

- Bias Score: discriminazioni o stereotipi.
- Safety Compliance: rispetto delle policy.

---

7. LLM-as-a-Judge

- Pairwise Comparison: confronto tra risposte.
- Valutazione Likert: punteggi da 1 a 5.
- Reward Models: allineamento a preferenze.

---

8. Metriche Operative

- Latency: tempo di risposta.
- Throughput: token/secondo.
- Cost per Token: costo operativo.
- Stability/Variance: coerenza delle risposte.