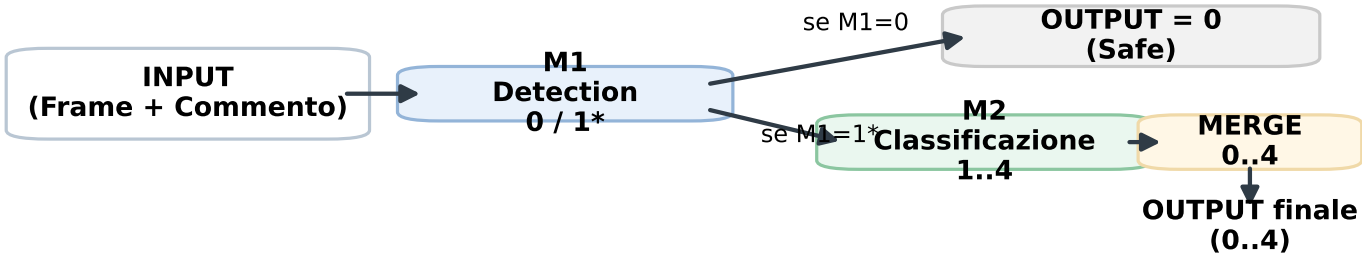
# TEST FINALE — Pipeline M1 (Detection) → M2 (Classificazione)

N test: 855 | Seed: [101, 285, 3692, 92] | Best seed (visual): 285
2026-02-17 23:00 | GPU: Tesla V100S-PCIE-32GB | Output: TEST_N3_PIPELINE_FULL

---

- M1: Detection binaria (0=Safe, 1*=Offensive dove 1*={1..4})
- M2: Classificazione multi-classe (training solo su campioni Offensive 1..4)
- Inference: M1 filtra; M2 gira solo sui campioni con M1=1*; merge nello spazio {0..4}
- Output finale: etichette {0,1,2,3,4} + metriche globali/per-classe + tempi + stabilità

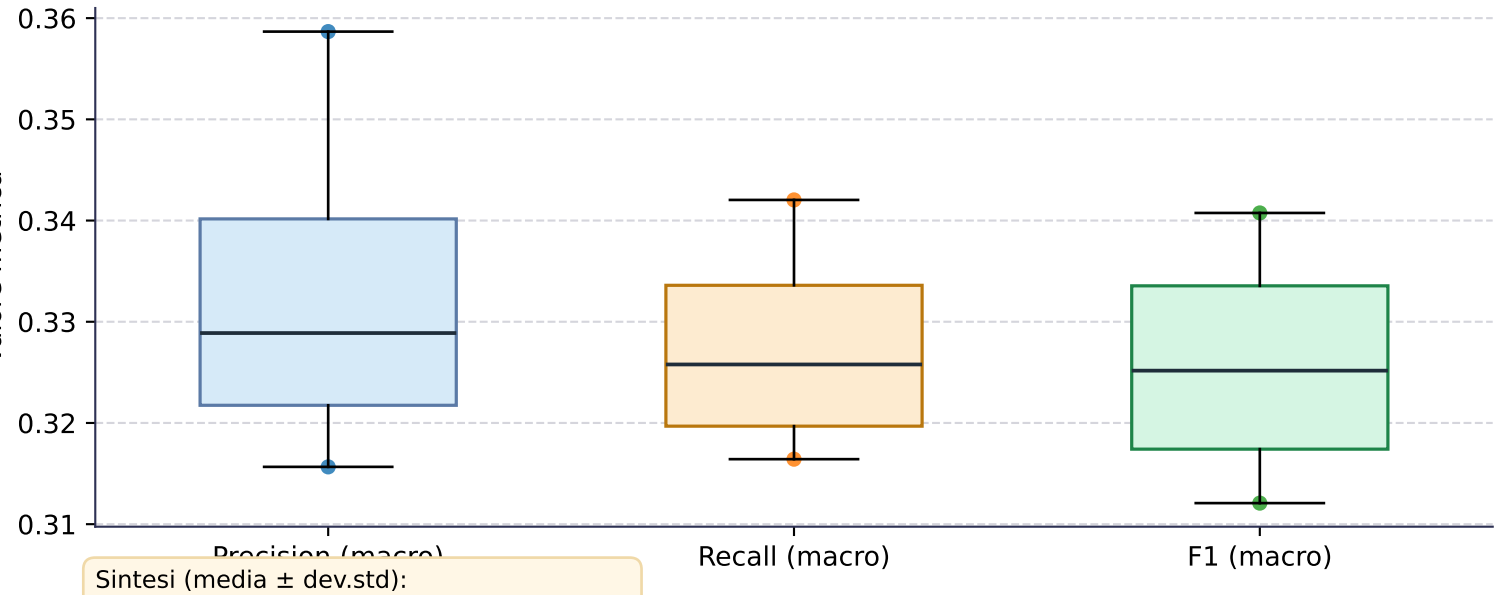



---

| Metrica | Baseline | Fine-Tuned (media ± dev.std) |
|---|---|---|
| Accuracy | 0.7520 | 0.7658 ± 0.0098 |
| Precision (macro) | 0.3705 | 0.3330 ± 0.0187 |
| Recall (macro) | 0.3327 | 0.3275 ± 0.0114 |
| F1 (macro) | 0.2909 | 0.3258 ± 0.0127 |

Nota: le metriche Fine-Tuned sono aggregate su seed (media ± dev.std).

# STABILITÀ, TEMPI E AUDIT OPERATIVO




Sintesi (media ± dev.std):
• Precision (macro): 0.3330 ± 0.0187 (n=4)
• Recall (macro): 0.3275 ± 0.0114 (n=4)
• F1 (macro): 0.3258 ± 0.0127 (n=4)

| Voce | Valore (media ± dev.std) | Note |
|---|---|---|
| Train M1 | 391.5 ± 2.3 min | n=4 |
| Train M2 | 150.3 ± 0.3 min | n=4 |
| Inferenza M1 | 0.769 ± 0.046 sec/campione | n=4 |
| Inferenza M2 | 0.705 ± 0.074 sec/campione | n=4 |

| Indicatore | Valore (media ± dev.std) | Descrizione |
|---|---|---|
| Attivazione M2 | 24.5% ± 1.4% | quota campioni inviati a M2 (m2_calls / N) |
| Fallback/Invalid | 0.0% ± 0.0% | quota chiamate M2 finite in fallback (fallback_events / m2_calls) |
| Missing immagini (su M2) | 0.0% ± 0.0% | quota chiamate M2 con immagini mancanti (m2_missing_images / m2_calls) |
| Missing pred M1 | 0.0% ± 0.0% | quota campioni senza predizione M1 (m1_missing / N) |

# DETTAGLIO PER CLASSE E MATRICE DI CONFUSIONE

Nota: tabella = media ± dev.std su seed; matrice = best seed (solo visual).

---

| Classe | Label | P (base) | P (FT) | R (base) | R (FT) | F1 (base) | F1 (FT) | ΔF1 |
|---|---|---|---|---|---|---|---|---|
| 0 | Safe | 0.778 | 0.873 ± 0.007 | 0.969 | 0.876 ± 0.014 | 0.863 | 0.875 ± 0.007 | +0.011 |
| 1 | Flaming | 0.304 | 0.201 ± 0.099 | 0.583 | 0.125 ± 0.048 | 0.400 | 0.153 ± 0.063 | -0.247 |
| 2 | Denigration | 0.588 | 0.497 ± 0.020 | 0.068 | 0.588 ± 0.034 | 0.121 | 0.538 ± 0.013 | +0.417 |
| 3 | Sexual | 0.182 | 0.093 ± 0.008 | 0.043 | 0.049 ± 0.011 | 0.070 | 0.063 ± 0.006 | -0.007 |
| 4 | Racism | 0.000 | 0.000 ± 0.000 | 0.000 | 0.000 ± 0.000 | 0.000 | 0.000 ± 0.000 | +0.000 |

---

**Reale \ Predetto**

| | Safe | Flaming | Denigr. | Sexual | Racism |
|---|---|---|---|---|---|
| **Safe** | **557** ✓ | 0 | 72 | 14 | 1 |
| **Flaming** | 5 | **2** ✓ | 5 | 0 | 0 |
| **Denigr.** | 45 | 3 | **94** ✓ | 5 | 1 |
| **Sexual** | 21 | 1 | 22 | **2** ✓ | 0 |
| **Racism** | 3 | 0 | 2 | 0 | **0** ✓ |

---

- Pipeline: M1 (0 vs 1*) → M2 (1..4) → merge (0..4).
- Fine-Tuned (media ± dev.std): Accuracy = 0.7658 ± 0.0098, F1 macro = 0.3258 ± 0.0127.
- Per-classe: confronto baseline vs fine-tuned (P/R/F1 accoppiati) + ΔF1.
- Instradamento: attivazione M2 = 24.5%; fallback/invalid su M2 = 0.0%.
- Matrice: ✓ sulla diagonale = predizioni corrette; numeri = conteggi reali.