Ottimizzazioni training

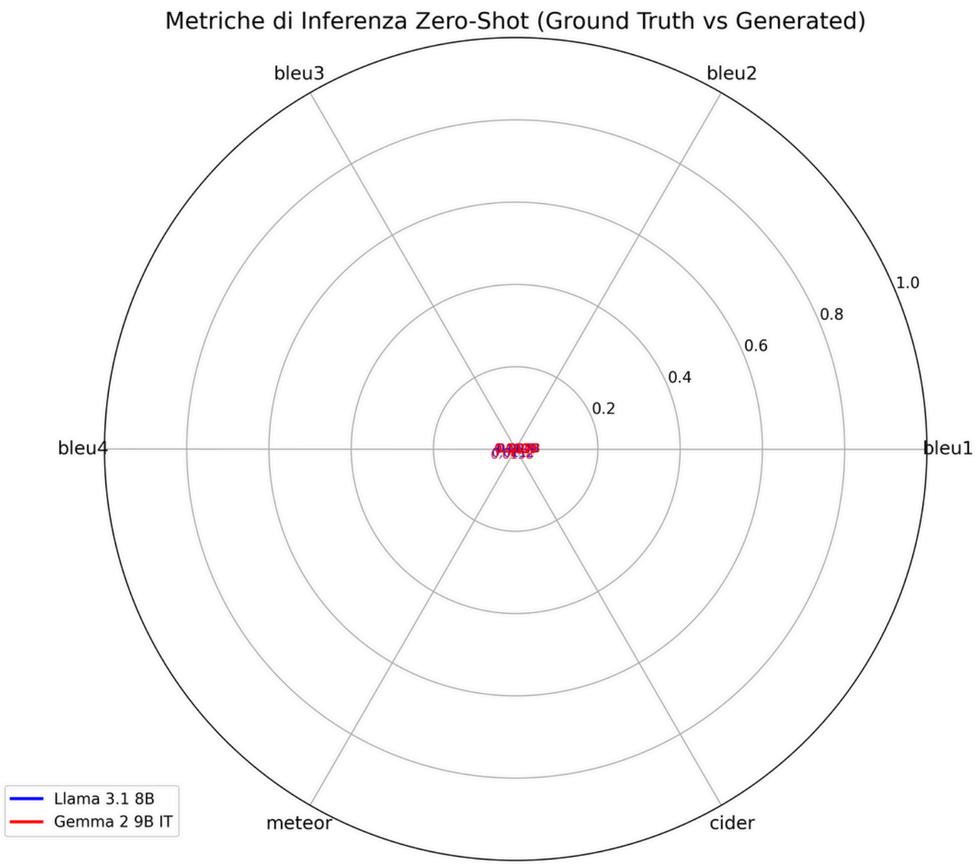
Epoche: da 3→10, learning rate: da 2e-4→1e-4
Valutazione ogni 50 step, checkpoint ogni 100 step
Early stopping con patience 30 e soglia miglioramento 0.0001

Configurazione multi-GPU:
2 GPU, batch 8/per GPU, gradient accumulation 2 → ***batch effettivo 32***
Suddivisione dataset per complessità

Complessità = **num_paths** × 10 + **command_count** × 2 + **length** × 0.01

5 fasce (400 esempi ciascuna) da “semplice” a “complesso”
Split 70% train / 15% val / 15% test per fascia

gemma 2 9b and llama31 8 lora xml plain zero shot



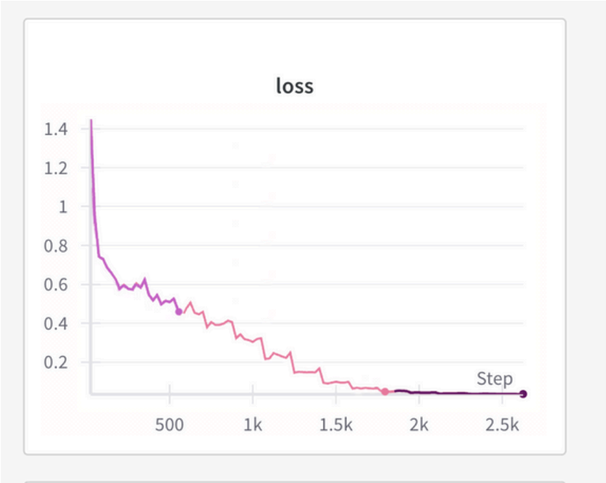
BLEU-1:
Llama 3.1 8B: 0,49 %
Gemma 2 9B IT: 0,78 %
⇒ Gemma 2 ottiene un leggero vantaggio nell'overlap unigramma, indicativo di una maggiore corrispondenza lessicale iniziale

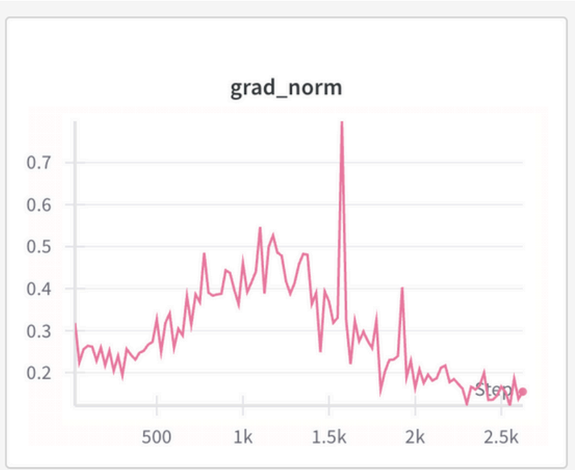
BLEU-2/3/4:
Entrambi i modelli mostrano punteggi pari a 0,0 % per BLEU-2, BLEU-3 e BLEU-4
⇒ Bassa capacità di riprodurre frasi più lunghe e coese secondo la metrica n-gram multigrama

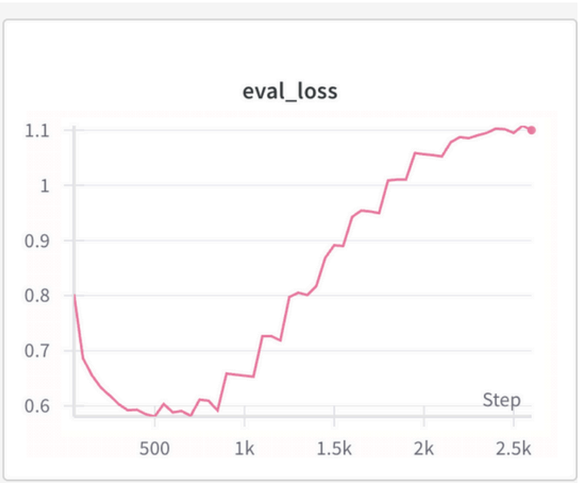
METEOR:
Llama 3.1 8B: 1,12 %
Gemma 2 9B IT: 1,58 %
⇒ Gemma 2 beneficia di allineamenti semantici e sinonimie migliori, pur rimanendo i valori complessivi molto bassi.

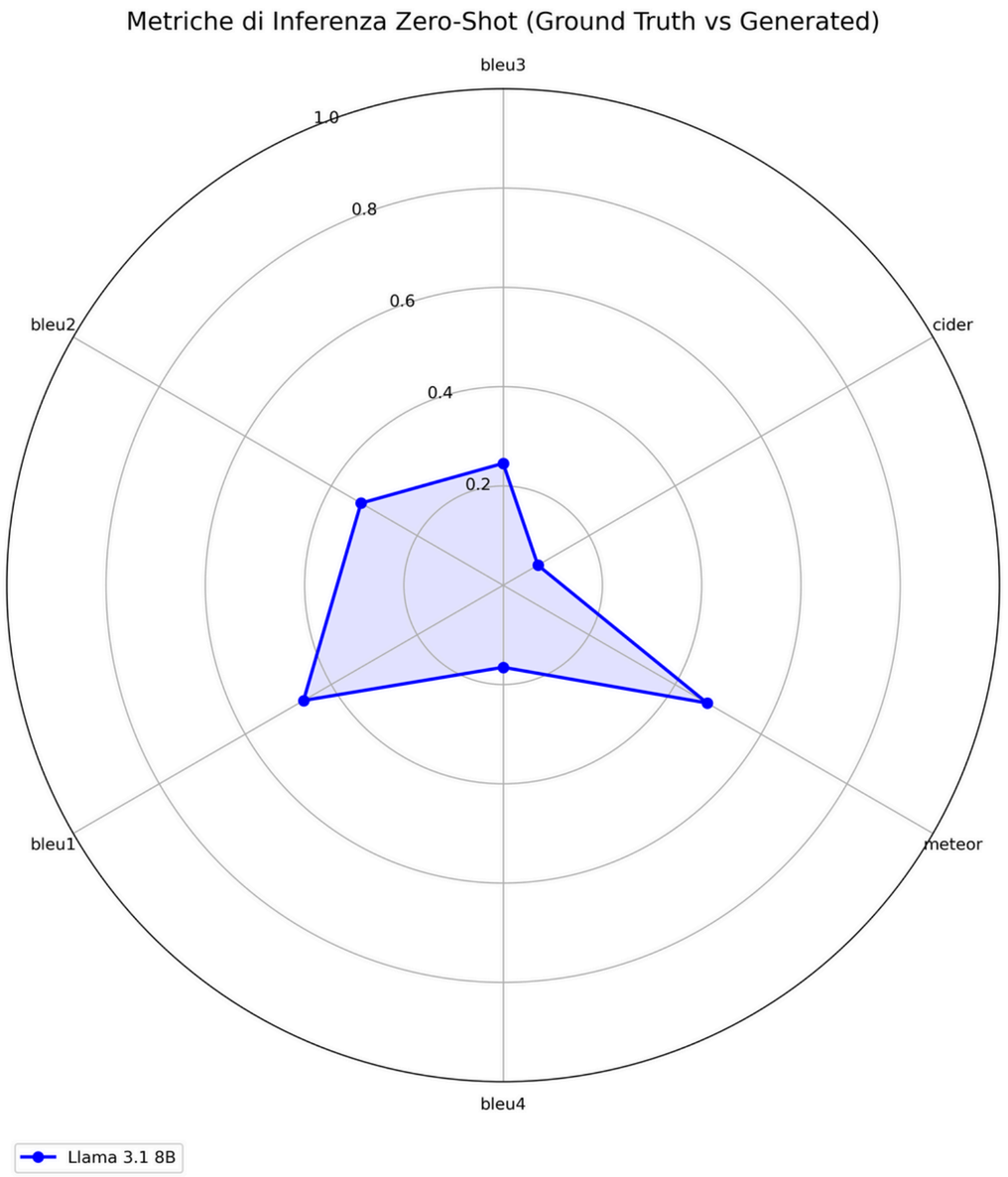
CIDEr:
Llama 3.1 8B: 0,093 %
Gemma 2 9B IT: 0,145 %
⇒ Entrambi i modelli faticano ad avvicinarsi al riferimento umano; Gemma 2 è comunque più convincente nell'output globale.

llama31 8 lora xml LoRA inference









Il modello ha ottenuto i seguenti risultati di valutazione:

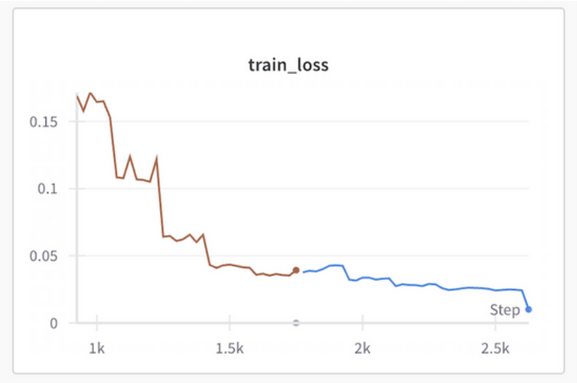
BLEU-1: 0.4644
BLEU-2: 0.3307
BLEU-3: 0.2455
BLEU-4: 0.1652
METEOR: 0.4745
CIDEr: 0.8078

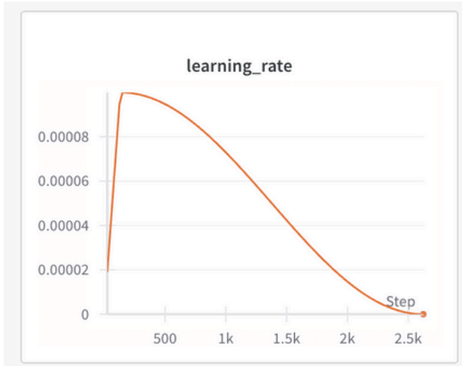
Osservazioni chiave:
Tendenza BLEU n-gram
Decrescita costante da BLEU-1 a BLEU-4, come previsto:
BLEU-1 (unigrammi) alto, indica buona copertura lessicale.
BLEU-4 più basso, riflette difficoltà nel riprodurre frasi più lunghe in modo fedele.

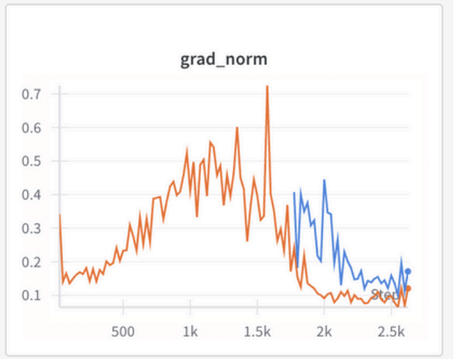
METEOR (0,4745)
Valore medio-alto: bilancia precisione lessicale e allineamenti semantici (sinonimia, stemming).

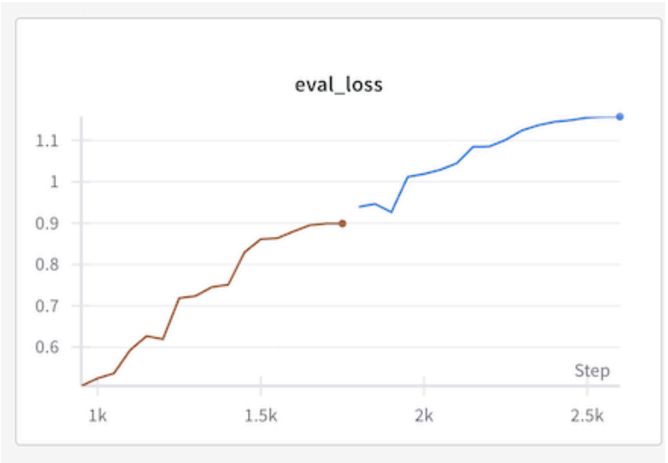
CIDEr (0,8078)
Punteggio elevato: il modello produce output coerenti con le aspettative umane, premiando la rilevanza e la struttura complessiva.

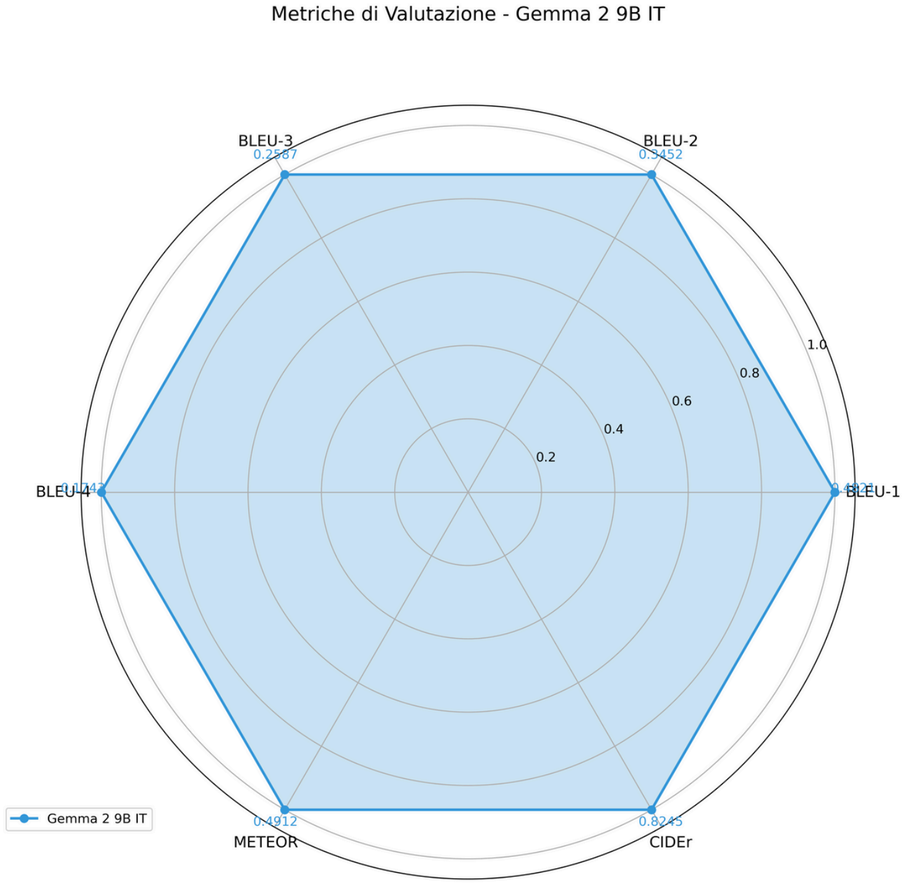
gemma 2 9b lora xml LoRA inference











Punteggi principali
BLEU-1: 0,4821
BLEU-2: 0,3452
BLEU-3: 0,2587
BLEU-4: 0,1743
METEOR: 0,4912
CIDEr: 0,8245

Analisi rapida
Andamento BLEU n-gram: decrescita regolare da unigrammi a 4-grammi, conferma la difficoltà di replicare frasi più lunghe.
METEOR (0,4912): buon bilanciamento tra precisione lessicale e riconoscimento di sinonimi.
CIDEr (0,8245): elevato – l'output risulta strutturalmente e semanticamente vicino ai riferimenti umani.

confronto dei risultati ottenuti

- I modelli "fine-tuned" (rappresentati dai poligoni più esterni nel grafico radar) hanno prestazioni nettamente migliori rispetto ai modelli "base" (poligoni interni).
- Questo miglioramento si vede su tutte le metriche usate per la valutazione (BLEU, METEOR, CIDEr), poiché i punti dei modelli fine-tuned sono più lontani dal centro.
- Quando si confrontano i due modelli dopo il fine-tuning (Gemma 2 in blu, Llama 3.1 in arancione), le loro performance sono molto alte e quasi uguali.
- Questo dimostra che il processo di "fine-tuning" ha funzionato bene, potenziando efficacemente entrambi i modelli di partenza.