

NanoSocrates: a very small language model

COLAPIETRA Antonio Pio
TAVILLA Simone

Deep Learning Project
Prof. Vito Walter Anelli

Obiettivo

L'obiettivo è sviluppare un singolo modello Transformer Encoder-Decoder specializzato nella traduzione bidirezionale tra testo non strutturato e dati RDF strutturati nel dominio dei film.

Il modello è progettato per quattro task principali:

- **Text-to-RDF:** Convertire un testo in linguaggio naturale (es. la trama di un film) in un insieme di triple RDF che ne catturino il significato semantico;
- **RDF-to-Text:** Generare una descrizione testuale coerente e leggibile a partire da un insieme di triple RDF strutturate;
- **Completamento di Triple:** Prevedere una componente mancante (soggetto, predicato o oggetto) all'interno di una singola tripla RDF, un compito analogo al "fill-in-the-blank";
- **Generazione Contestuale di Triple:** Prevedere triple RDF successive basandosi su un contesto di triple già fornite, simulando un compito di completamento di grafi di conoscenza;

Pipeline

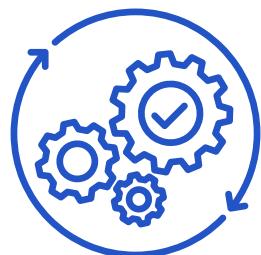


Pipeline

Modello ed addestramento



Implementazione della strategia a 3 fasi (Pre-training, Decoder Tuning, Full Fine-tuning) per un apprendimento graduale e stabile.



Valutazione e analisi



Esecuzione di una pipeline di validazione multi-metrica (BLEU, ROUGE, F1-Score per triple/entità, Accuracy) per una valutazione completa delle performance su ogni task



Addestramento

1.

Pre-training:
auto-supervisionato con
Span Corruption

2.

Decoder Tuning:
con encoder congelato per
un adattamento efficiente

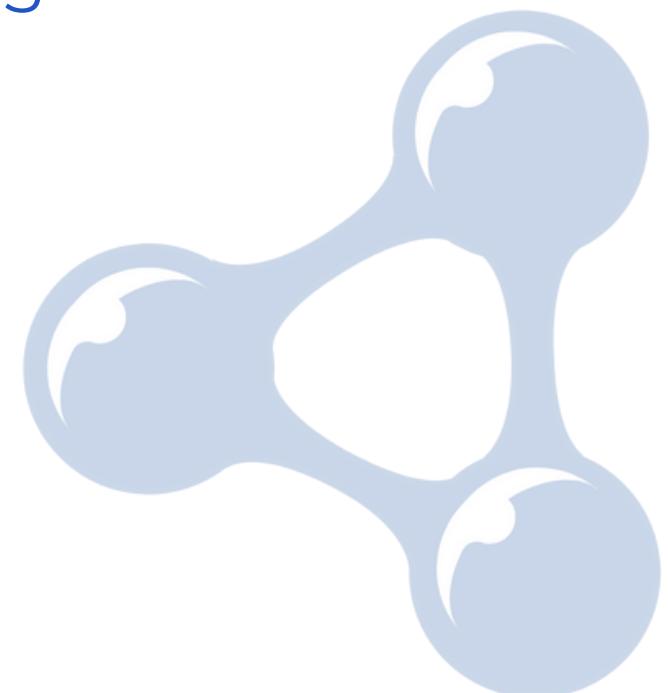
3.

Full Fine-Tuning:
a basso learning rate per la
massima specializzazione

Raccolta dei dati

La sfida principale era estrarre in modo efficiente decine di migliaia di record da un endpoint pubblico.

Per superare questa limitazione, è stata implementata una **query SPARQL ottimizzata** che utilizza una strategia di paginazione per scaricare i dati in blocchi, riducendo drasticamente il numero di richieste. La qualità è stata garantita tramite il whitelisting di predicati rilevanti per il dominio cinematografico e un filtro sulla lingua inglese.



Pre-Processing

pre-train

Prima di specializzare il modello è fondamentale insegnargli il "linguaggio" del nostro dominio.

Input:
file JSON dove ogni film è
rappresentato dal suo
abstract e da una lista di
triple RDF



Output:

corpus di pre-train unifica sia
gli abstract in linguaggio
naturale sia le triple RDF
linearizzate in formato
testuale

Questo corpus è il risultato di un processo di pulizia che include il filtraggio per lunghezza e la rimozione dei duplicati. La scelta metodologica più importante è stata il bilanciamento del corpus, assicurando una rappresentazione paritetica (circa 50/50) tra testo e dati strutturati.

Pre-Processing

fine-tuning

Serve un dataset supervisionato che insegni al modello a eseguire i suoi quattro compiti specifici.

Input:
file JSON dove ogni film è
rappresentato dal suo
abstract e da una lista di
triple RDF



Output:
Due file di testo paralleli
contenenti le coppie input-
output bilanciate e formattate
per l'addestramento
supervisionato del modello

Poiché questo processo genera un numero disomogeneo di esempi per task, è stata applicata una strategia di bilanciamento finale basata su percentuali fisse, assicurando che il modello riceva un'esposizione controllata a ogni abilità durante l'addestramento.

Tokenizzazione

Serve un dataset supervisionato che insegni al modello a eseguire i suoi quattro compiti specifici. Poiché questo processo genera un numero disomogeneo di esempi per task, è stata applicata una strategia di bilanciamento finale basata su percentuali fisse, assicurando che il modello riceva un'esposizione controllata a ogni abilità durante l'addestramento.

Input:

Il corpus di pre-training
unificato contenente testo
bilanciato tra abstract in
linguaggio naturale e triple
RDF linearizzate.



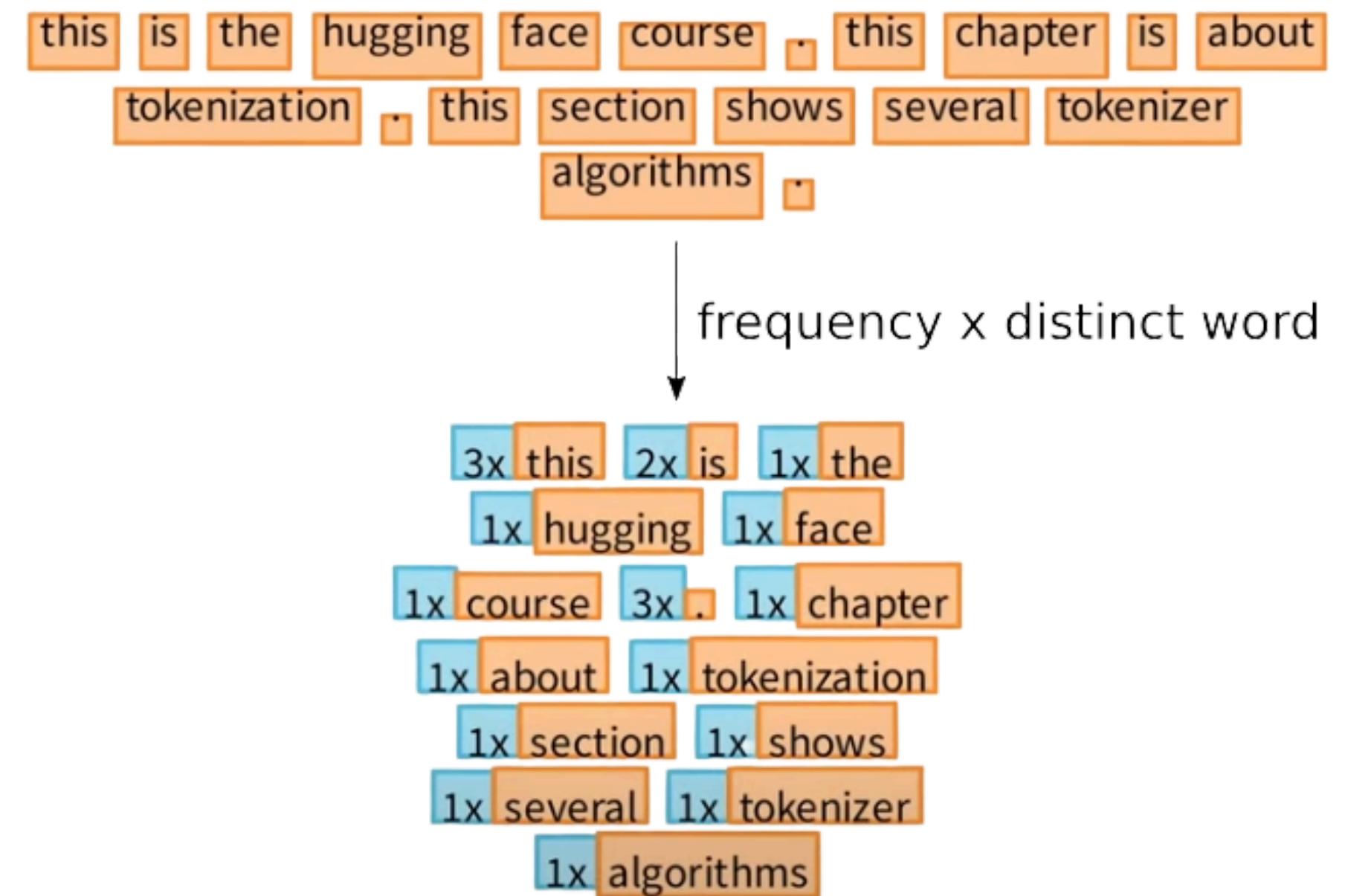
Output:

Un singolo file di configurazione
JSON che incapsula l'intero
vocabolario e le regole di
segmentazione, pronto per
essere utilizzato per convertire
qualsiasi testo del dominio in
una sequenza di token ID.

La natura ibrida dei nostri dati rende inadeguati i comuni tokenizer pre-addestrati, abbiamo quindi addestrato da zero un tokenizer Byte-Pair Encoding (BPE), utilizzando il corpus di pre-training bilanciato.

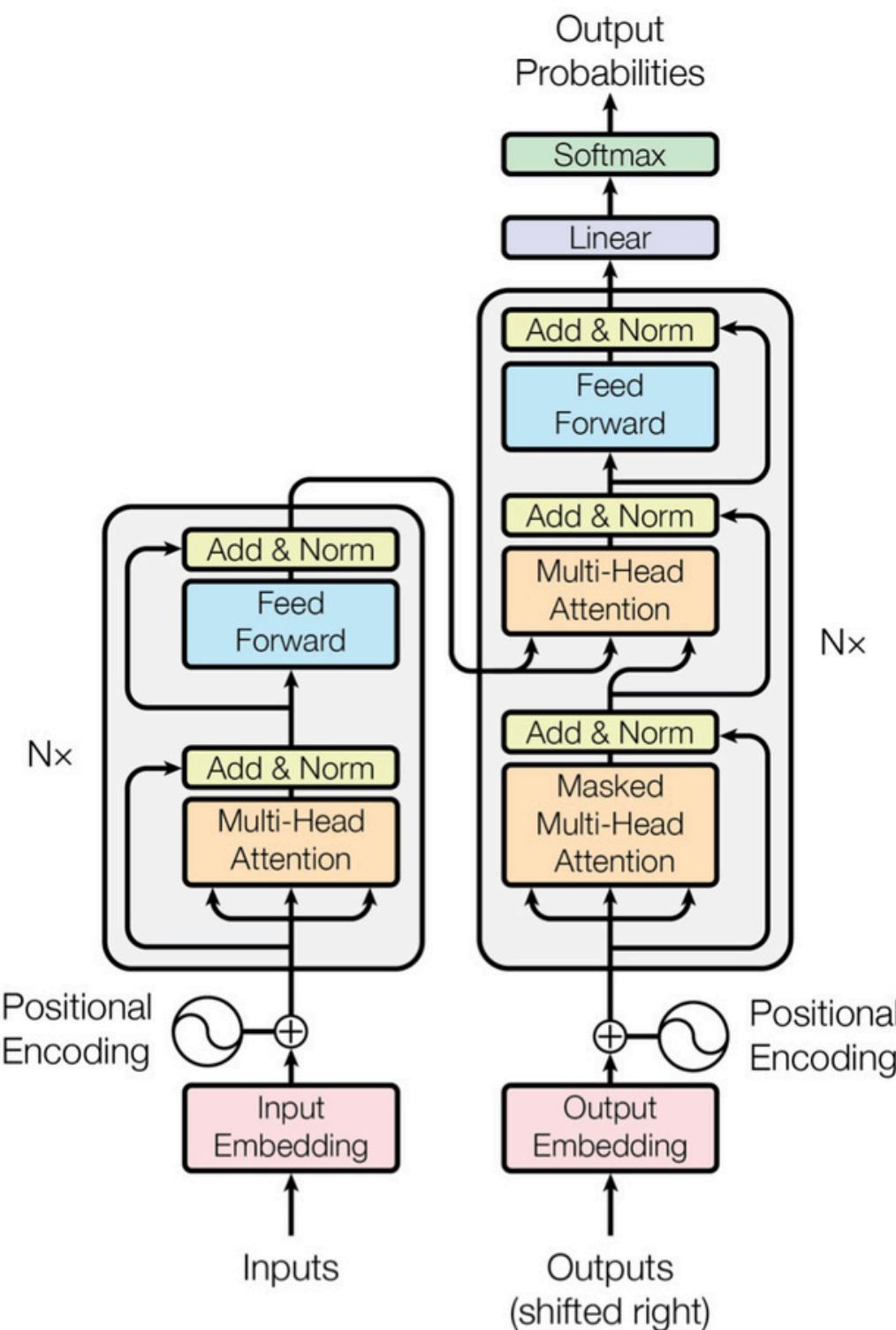
Tokenizzazione

Il vero cuore della nostra soluzione risiede nella configurazione del pre-tokenizer. Questo componente è stato istruito per isolare e trattare tutti i token speciali e la sintassi RDF, come i prefissi “dbr:” o i marcatori “<SUBJ>”, come unità atomiche e indivisibili, impedendone la frammentazione!



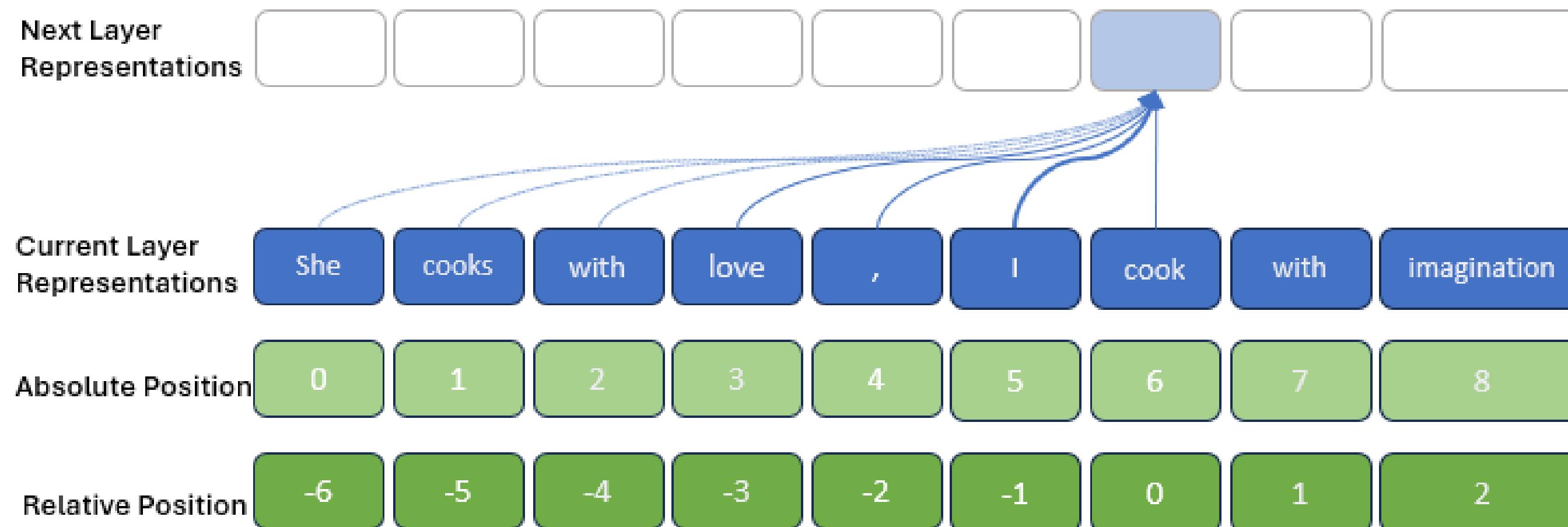
Modello

T5 style



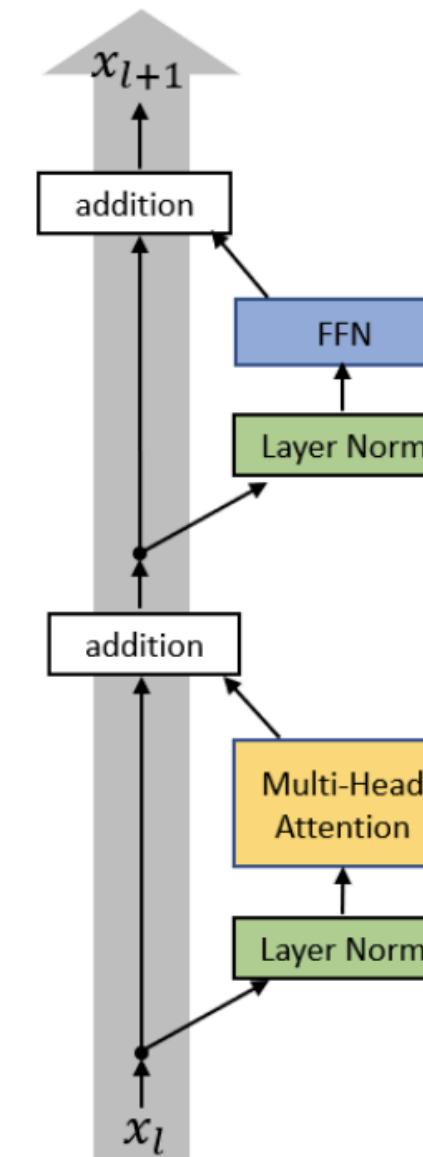
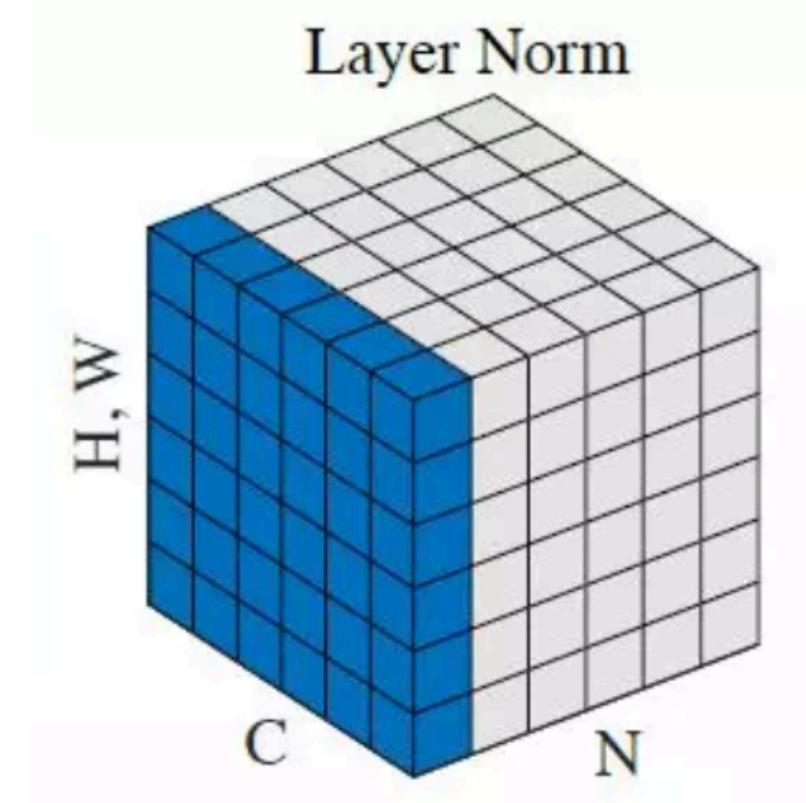
Modello

Relative Positional Bias



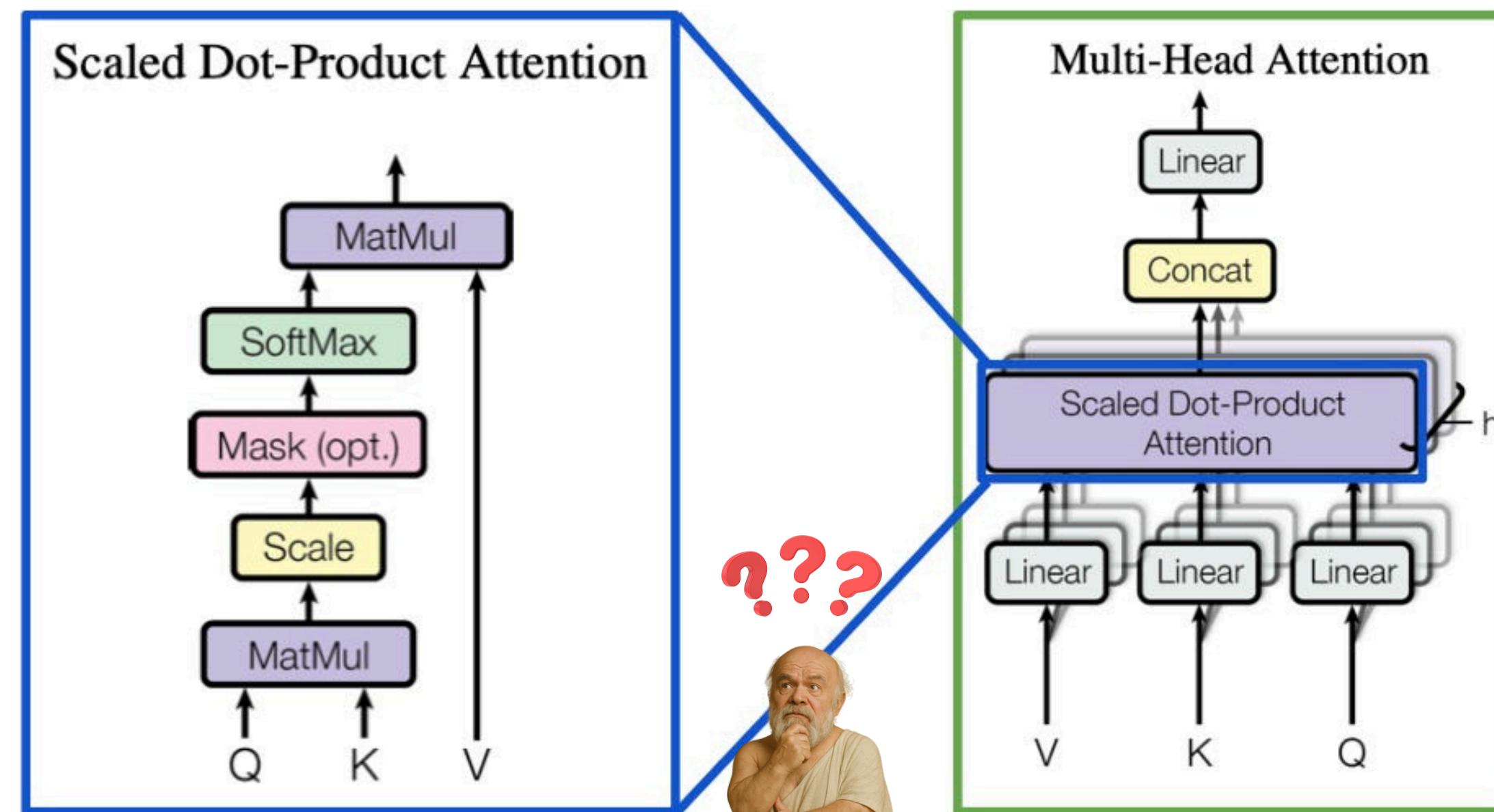
Modello

Layer Normalization



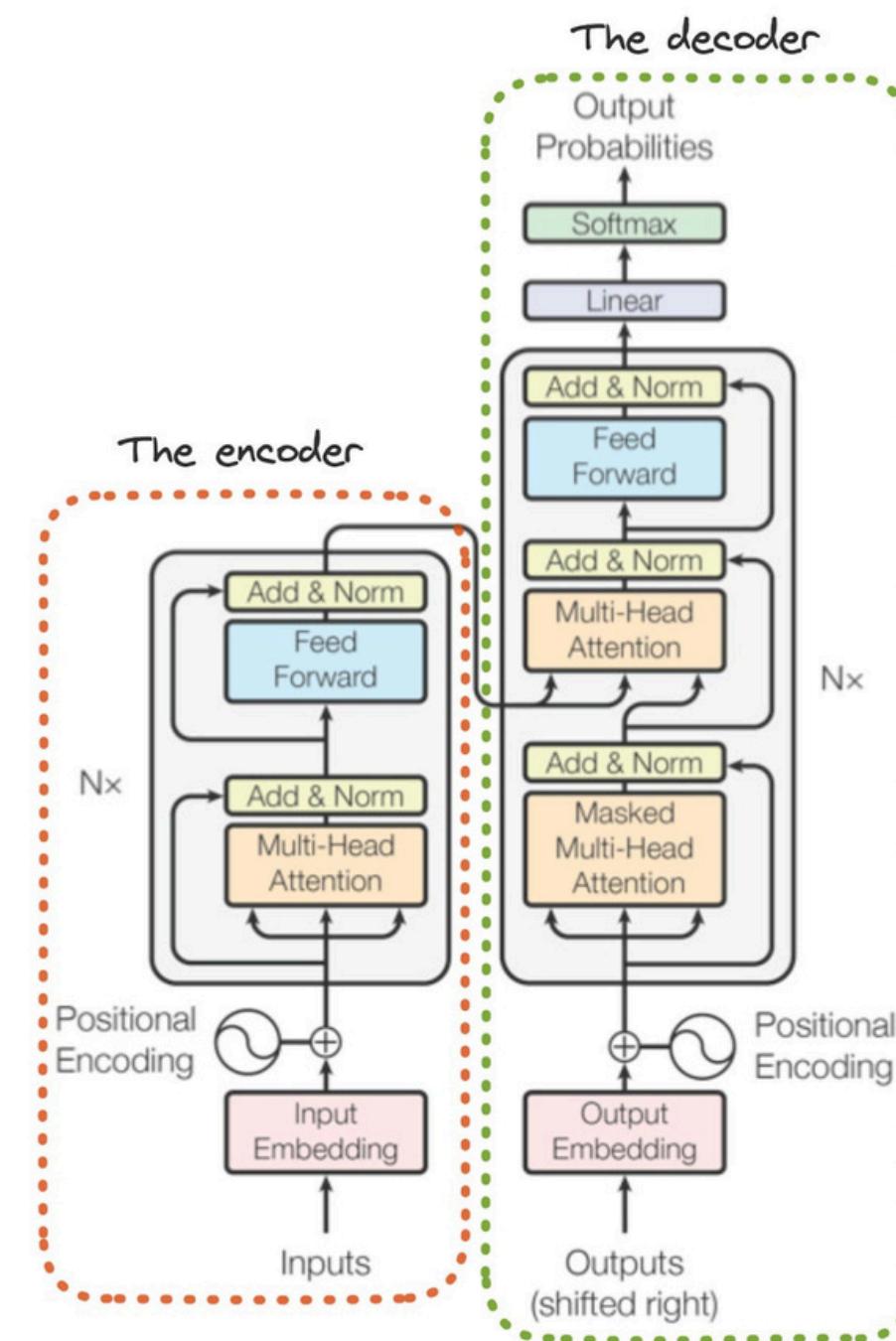
Modello

Meccanismo di Attenzione



Modello

Encoder/Decoder



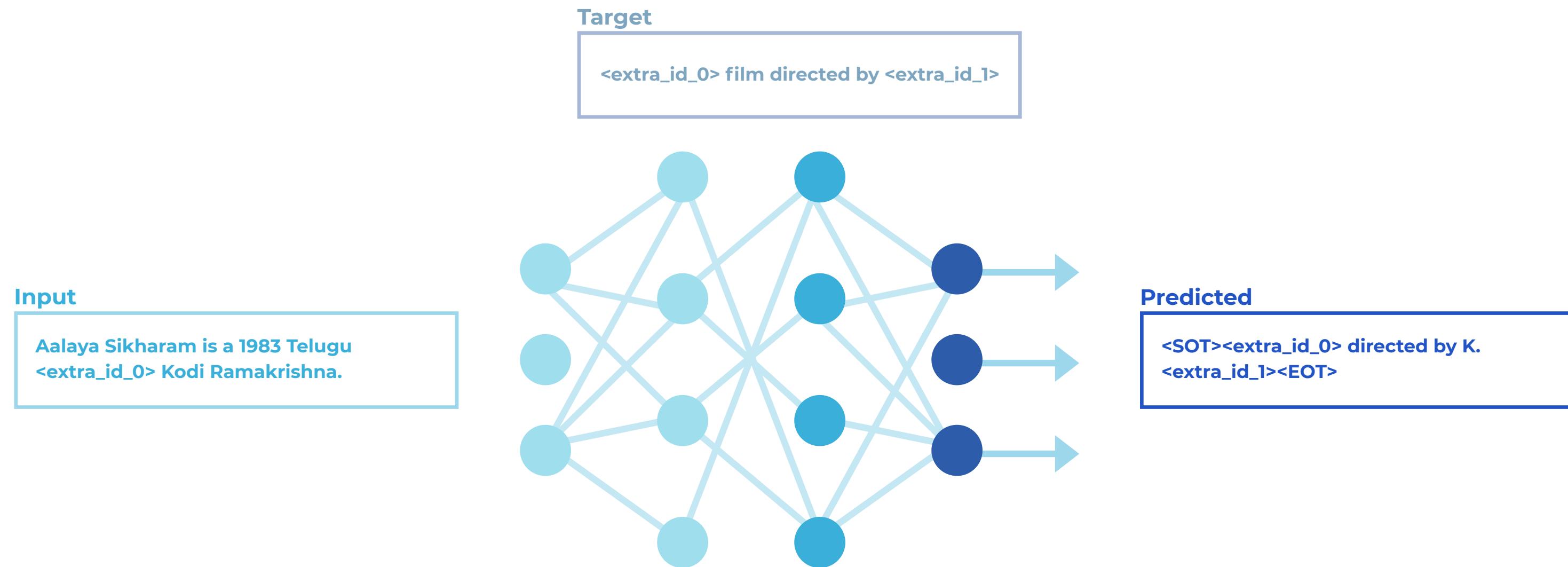
Sfide dell'addestramento

Model Collapse

Catastrophic Oblivion

Addestramento

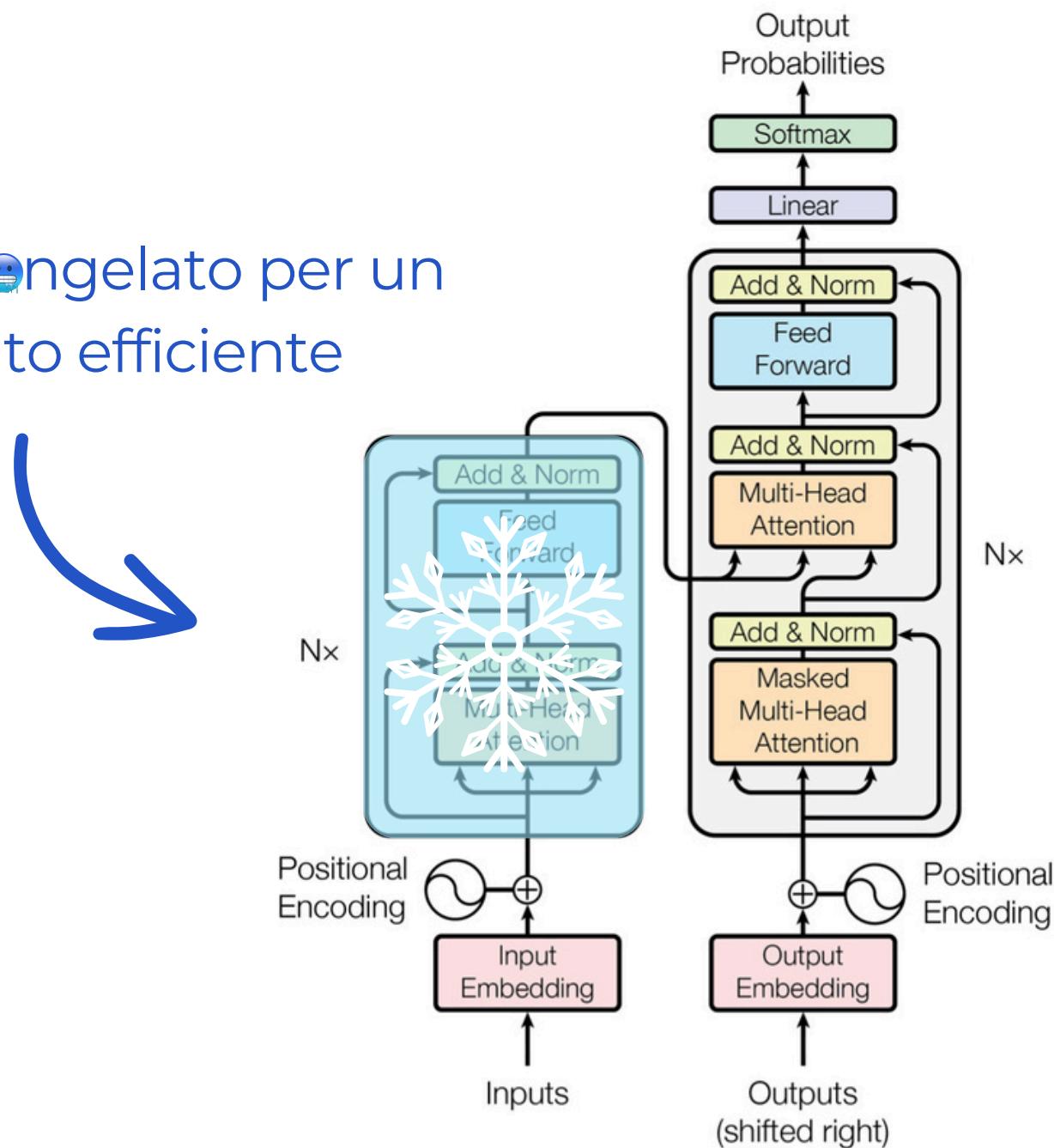
Pre-training



Addestramento

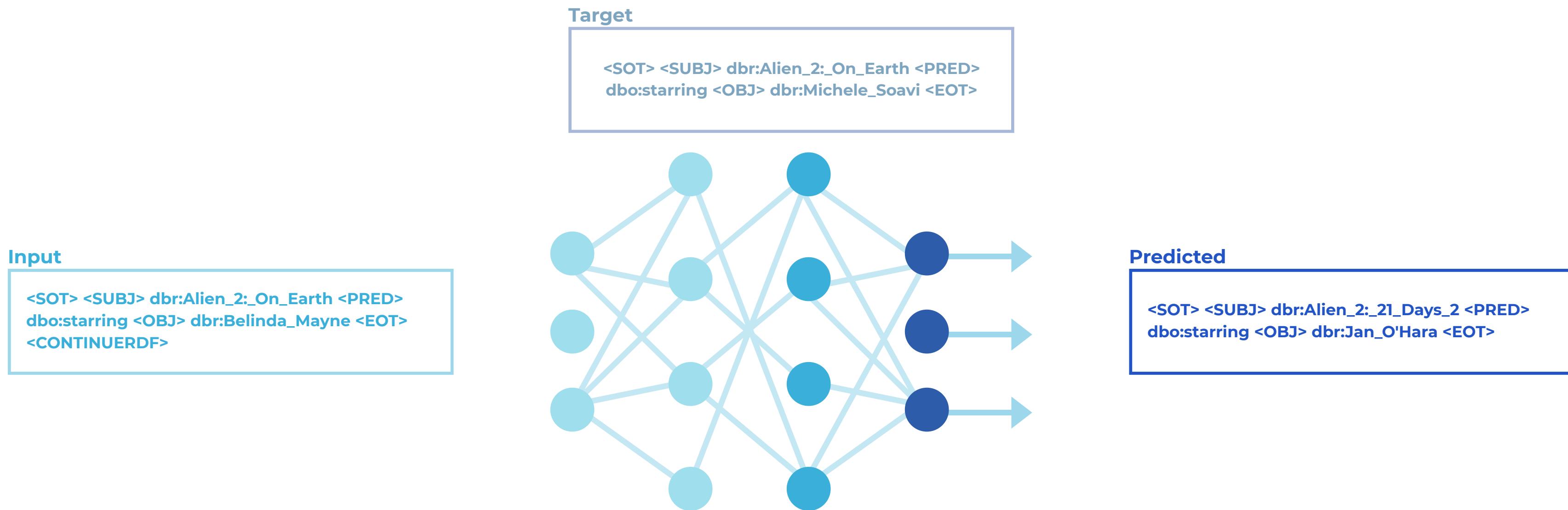
Decoder-tune

con encoder congelato per un
adattamento efficiente



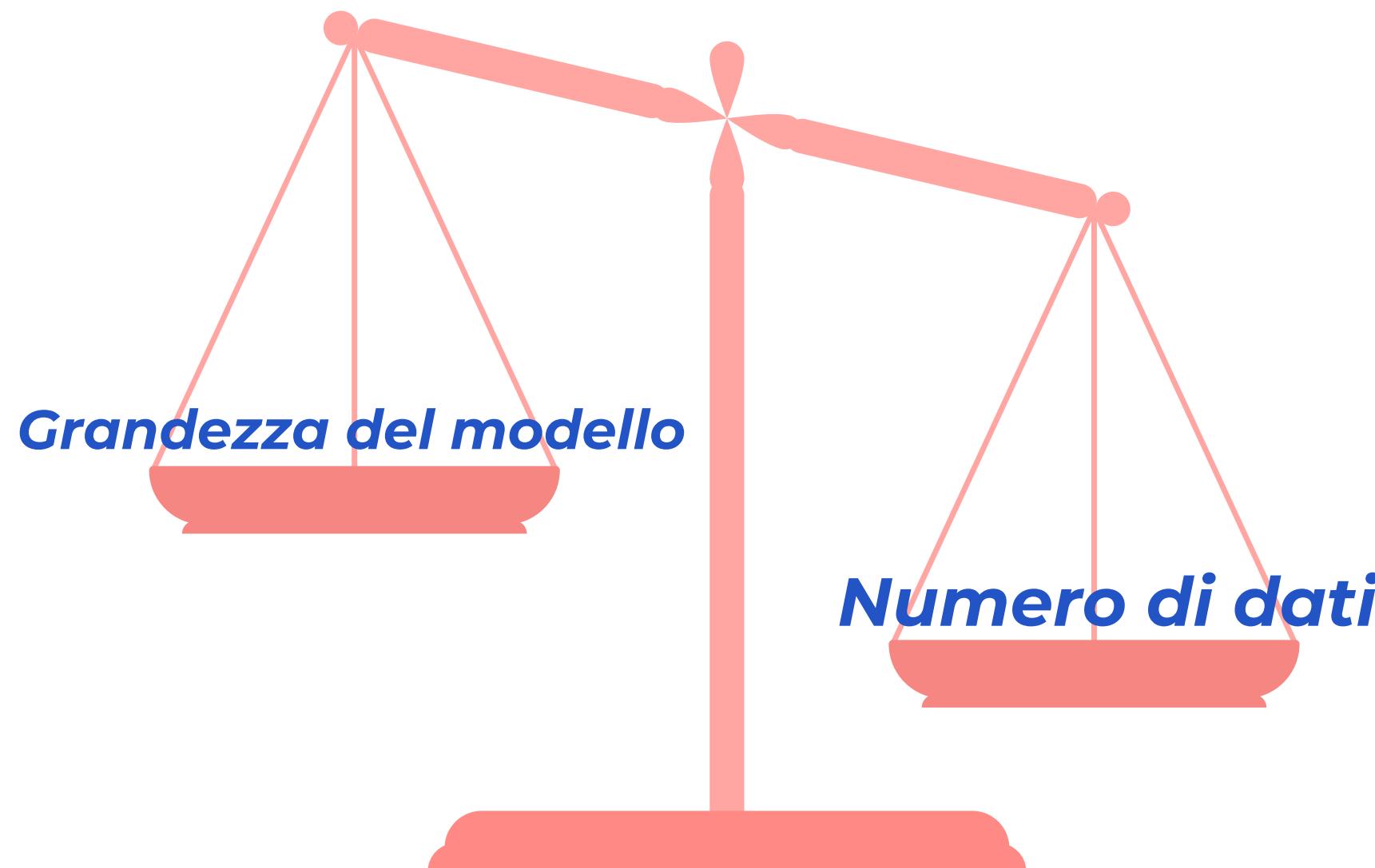
Addestramento

Full fine-tune

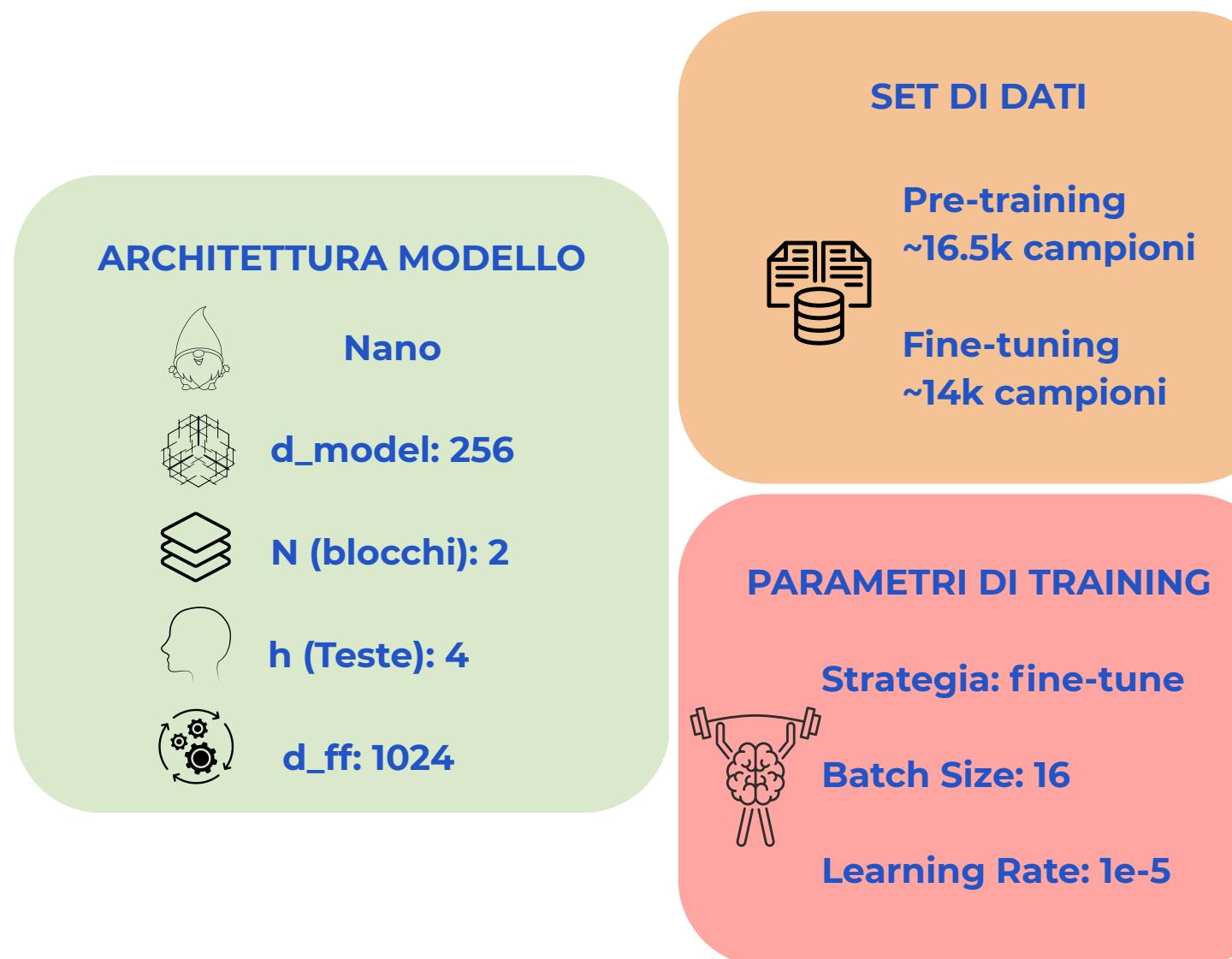


La nostra strategia

A causa delle risorse computazionali limitate, non abbiamo potuto testare direttamente la configurazione ottimale del nostro modello su un dataset enorme. Abbiamo quindi adottato una strategia per dimostrare la nostra tesi principale in modo indiretto.



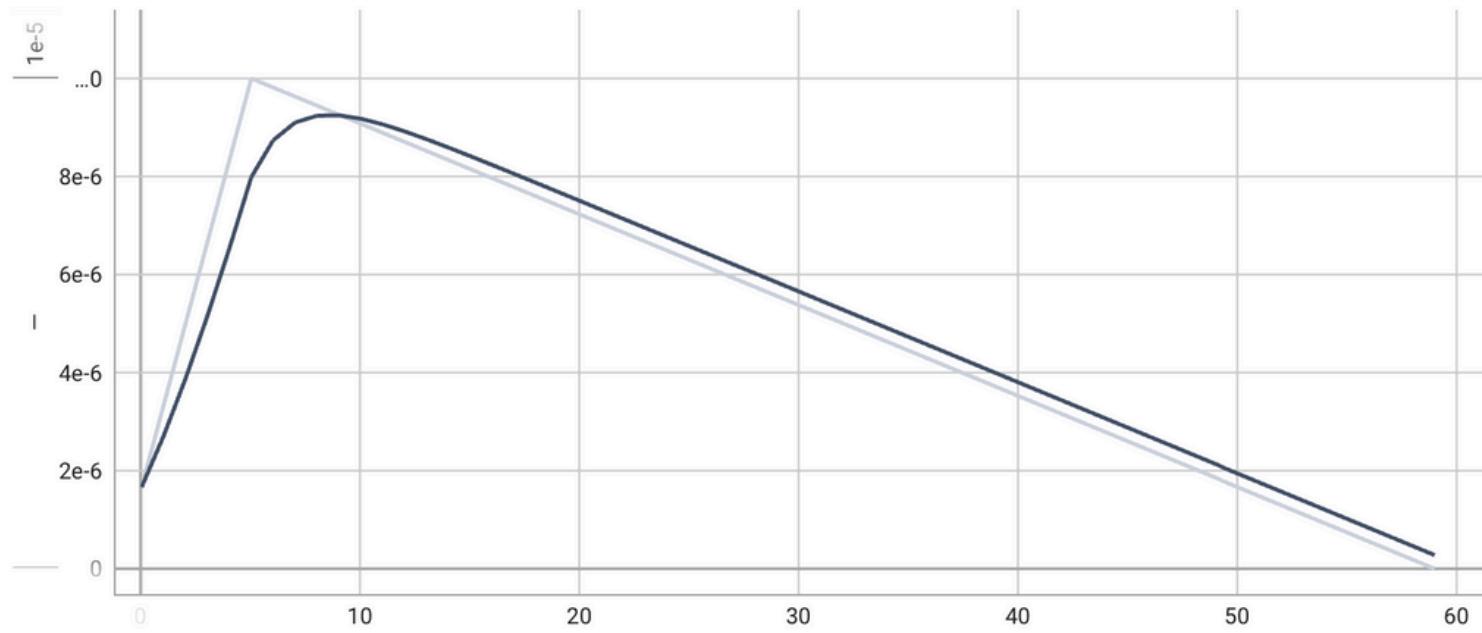
Esperimento 1



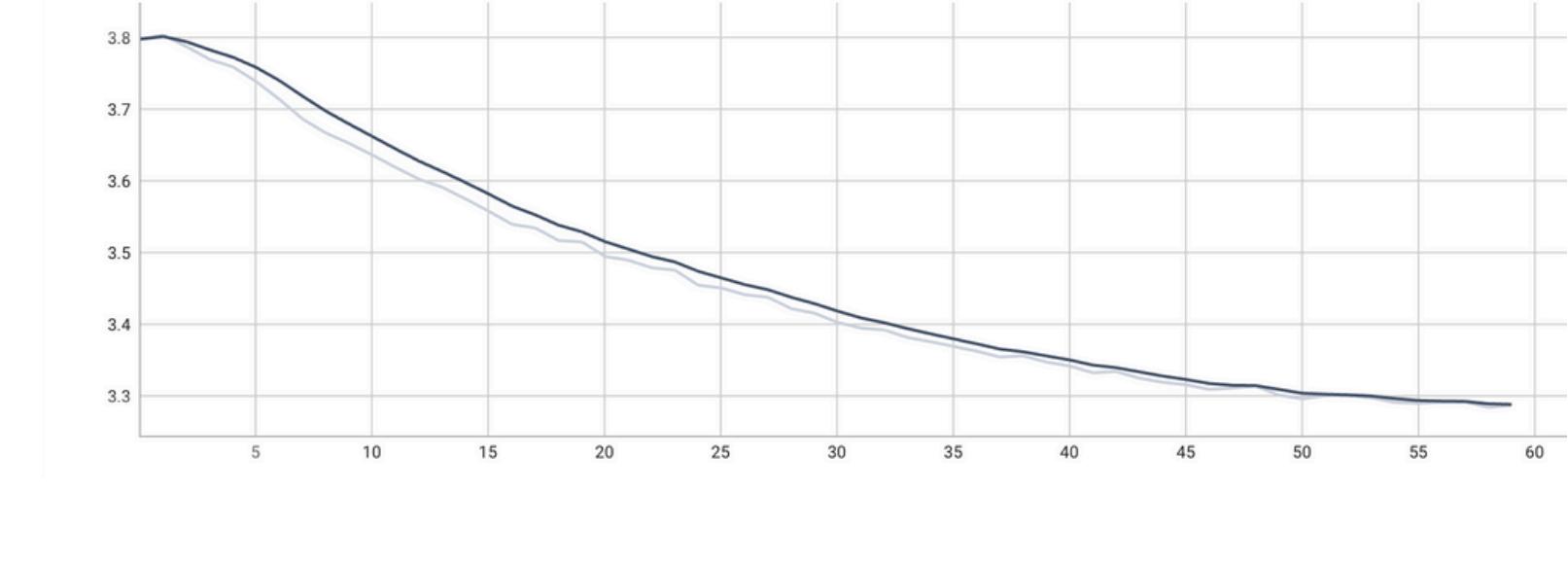
Categoria	Metrica / Componente	Precision	Recall	F1-Score
<i>Generale</i>	Validation Loss			2.2295
<i>NLG (Task RDF2Text)</i>	BLEU			0.0203
	METEOR			0.1506
	ROUGE-L			0.2157
<i>RDF a Livello di Entità</i>	Subjects	0.0367	0.0101	0.0158
	Predicates	0.9223	0.2536	0.3978
	Objects	0.1257	0.0346	0.0542
<i>MLM</i>	Accuracy (Soft)			0.2302

Esperimento 1

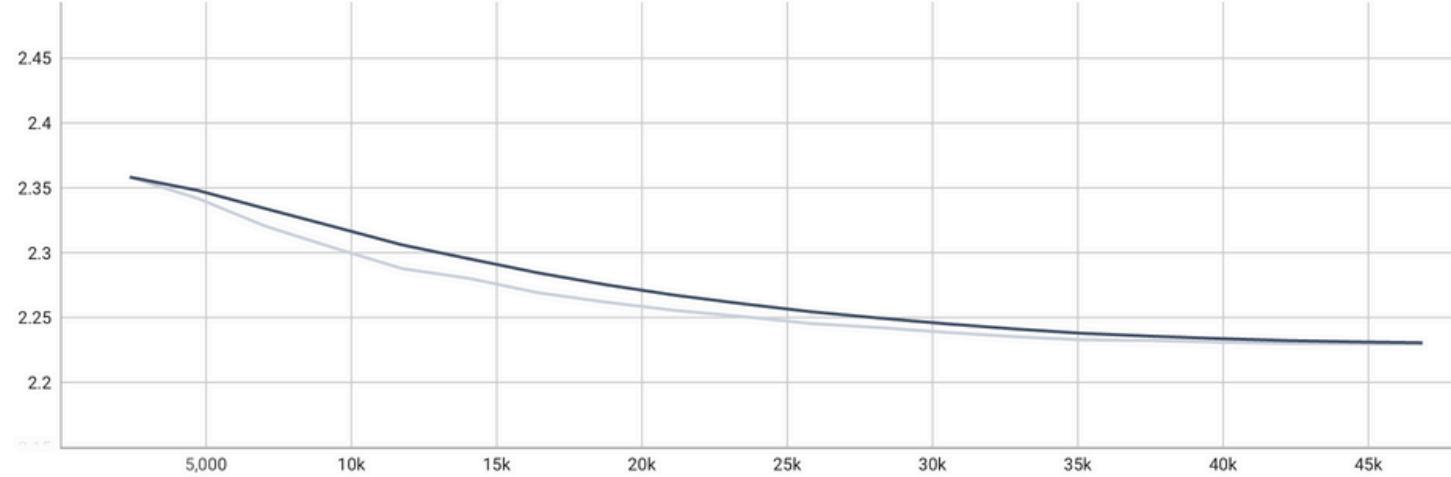
learning rate/full_finetune



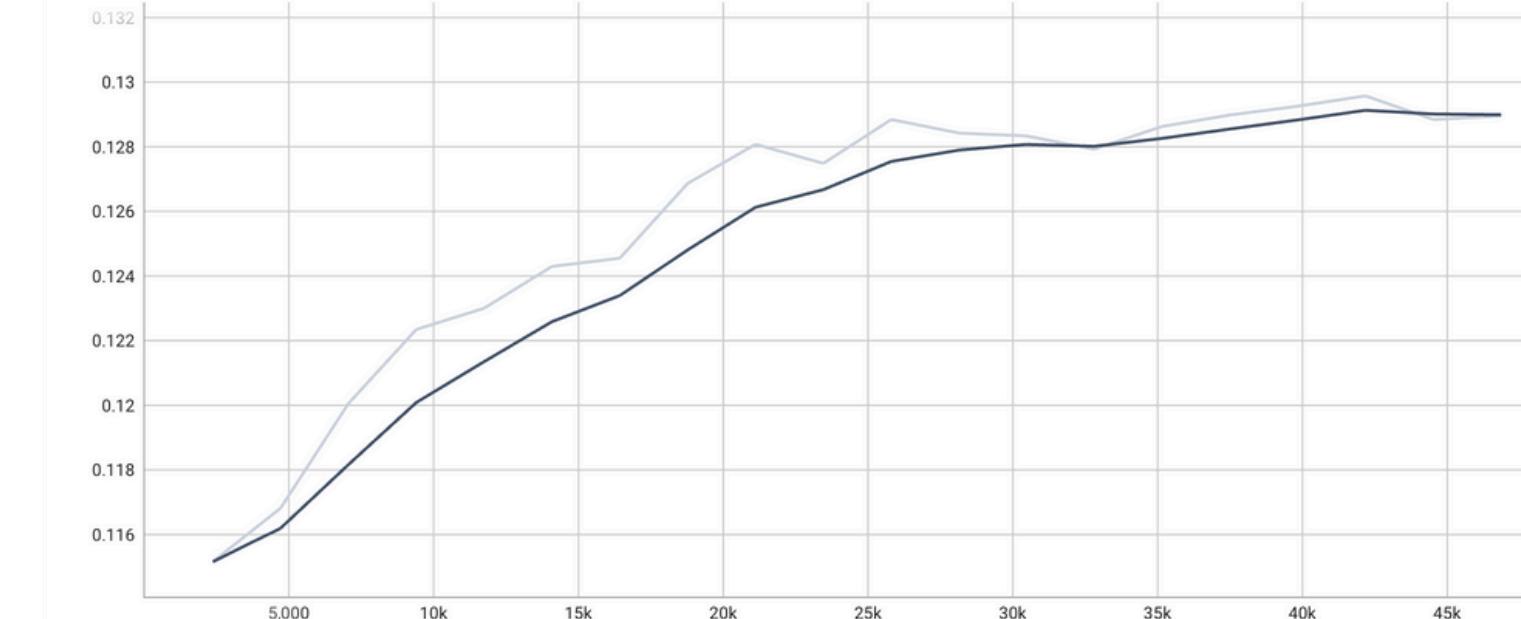
training_loss/full_finetune



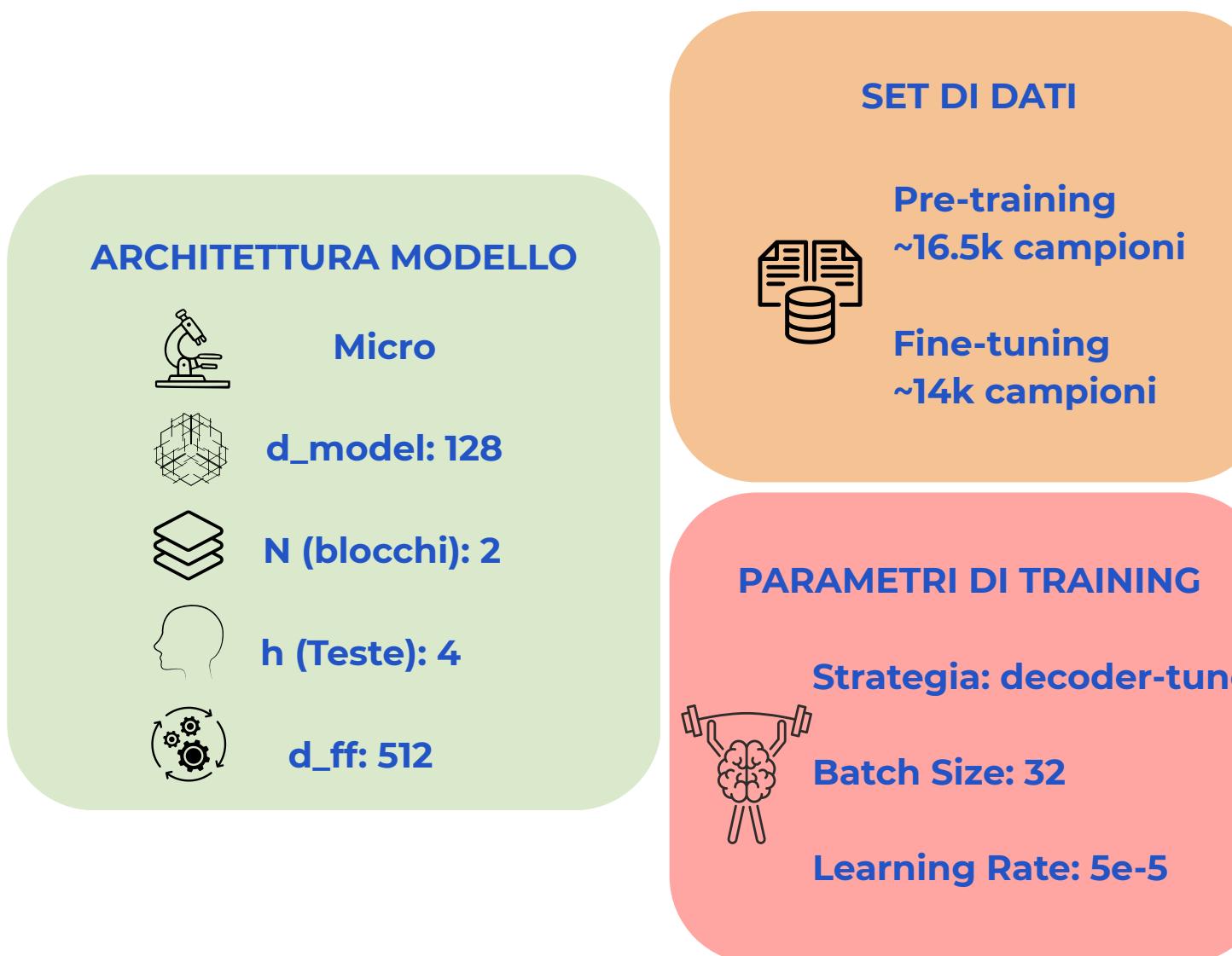
validation_loss/full_finetune



token_f1/full_finetune



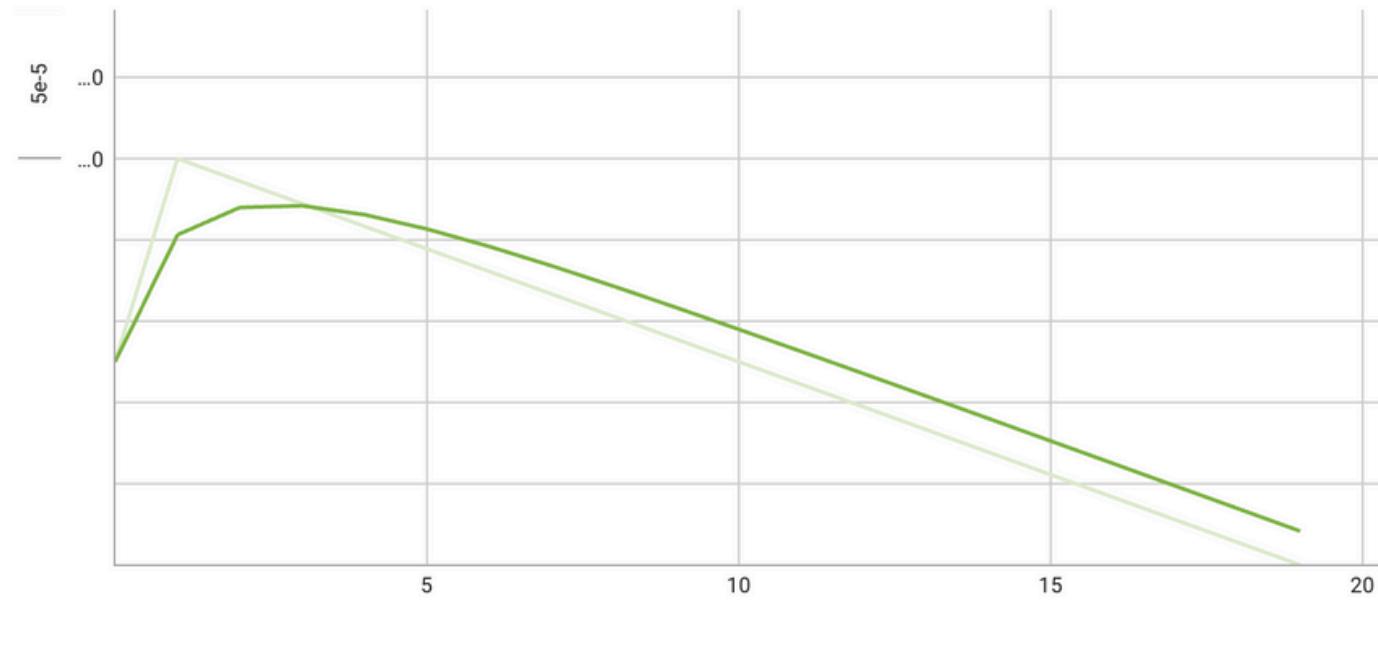
Esperimento 2



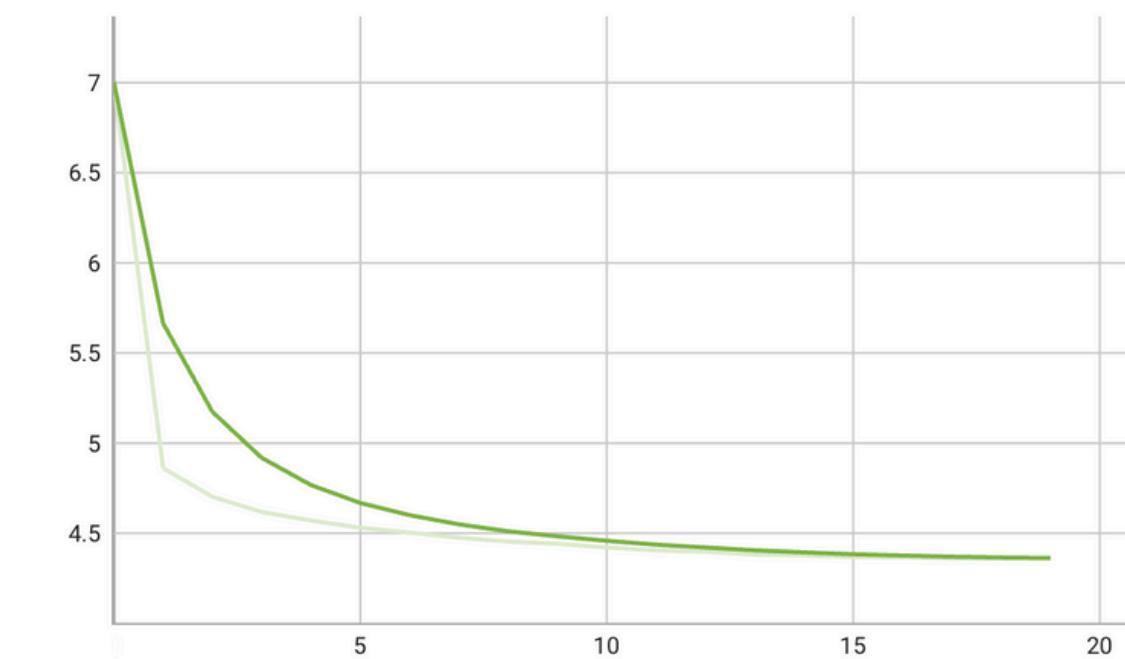
Categoria	Metrica / Componente	Precision	Recall	F1-Score
<i>Generale</i>	Validation Loss		3.0845	
<i>NLG (Task RDF2Text)</i>	BLEU		0.0208	
	METEOR		0.1082	
	ROUGE-L		0.1144	
<i>RDF a Livello di Entità</i>	Subjects	0.0000	0.0000	0.0000
	Predicates	1.0000	0.2443	0.3927
	Objects	0.0000	0.0000	0.0000
<i>MLM</i>	Accuracy (Soft)		0.0000	

Esperimento 2

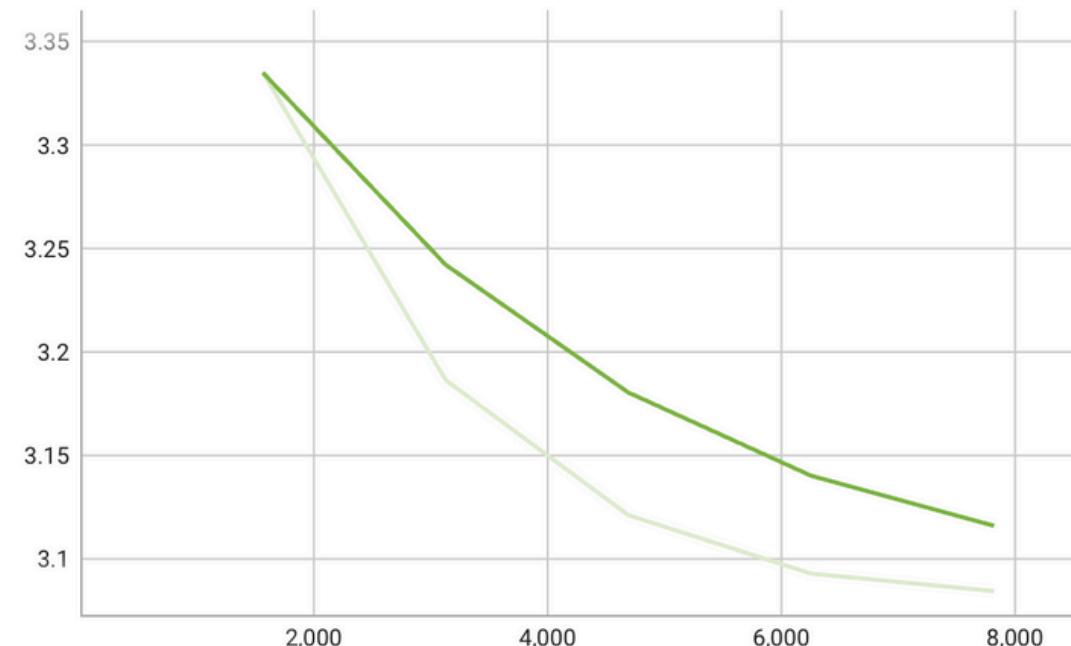
learning rate/decoder_tune



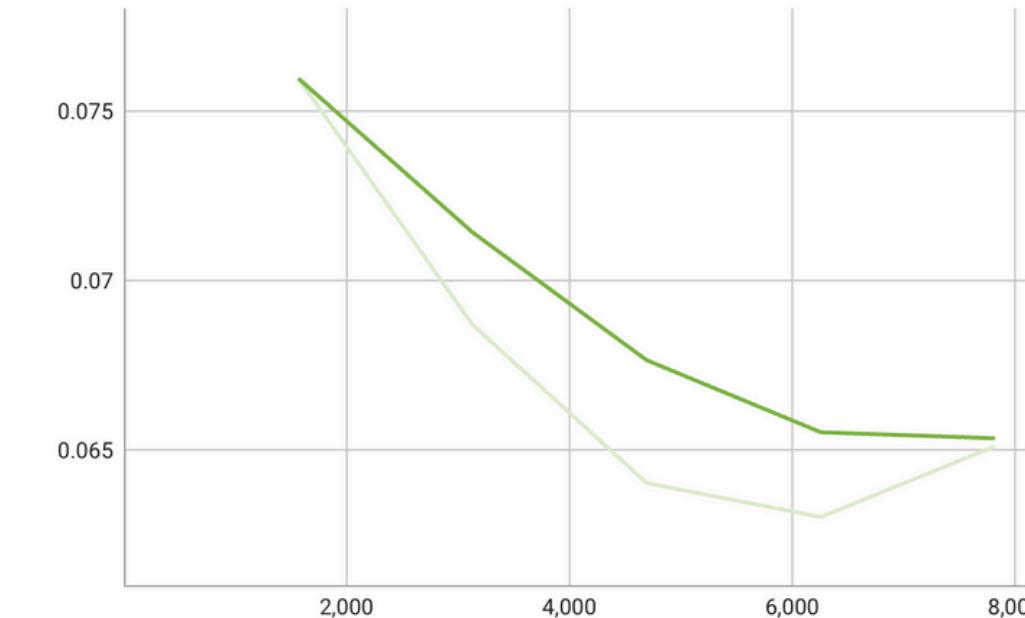
training_loss/decoder_tune



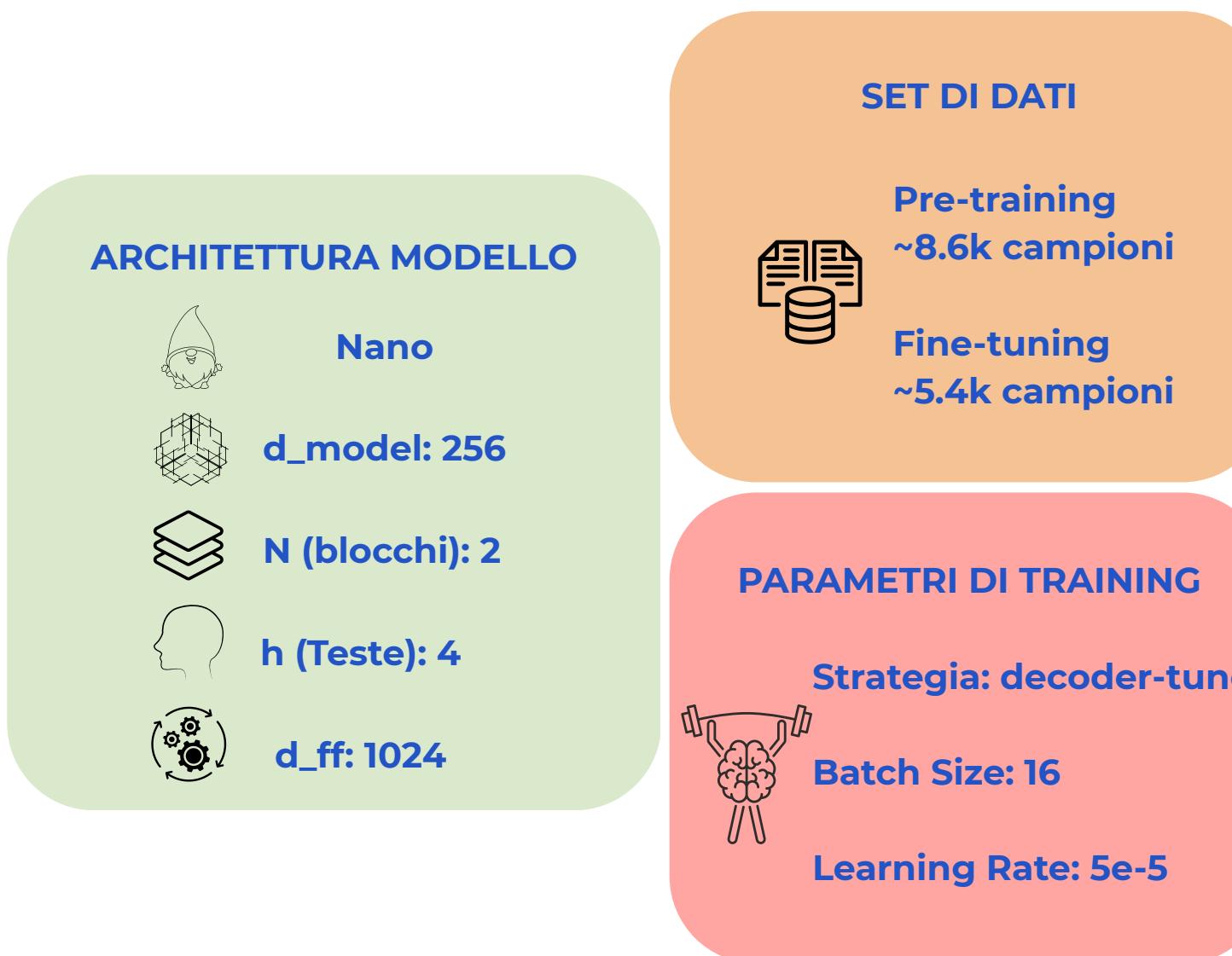
validation_loss/decoder_tune



token_f1/decoder_tune



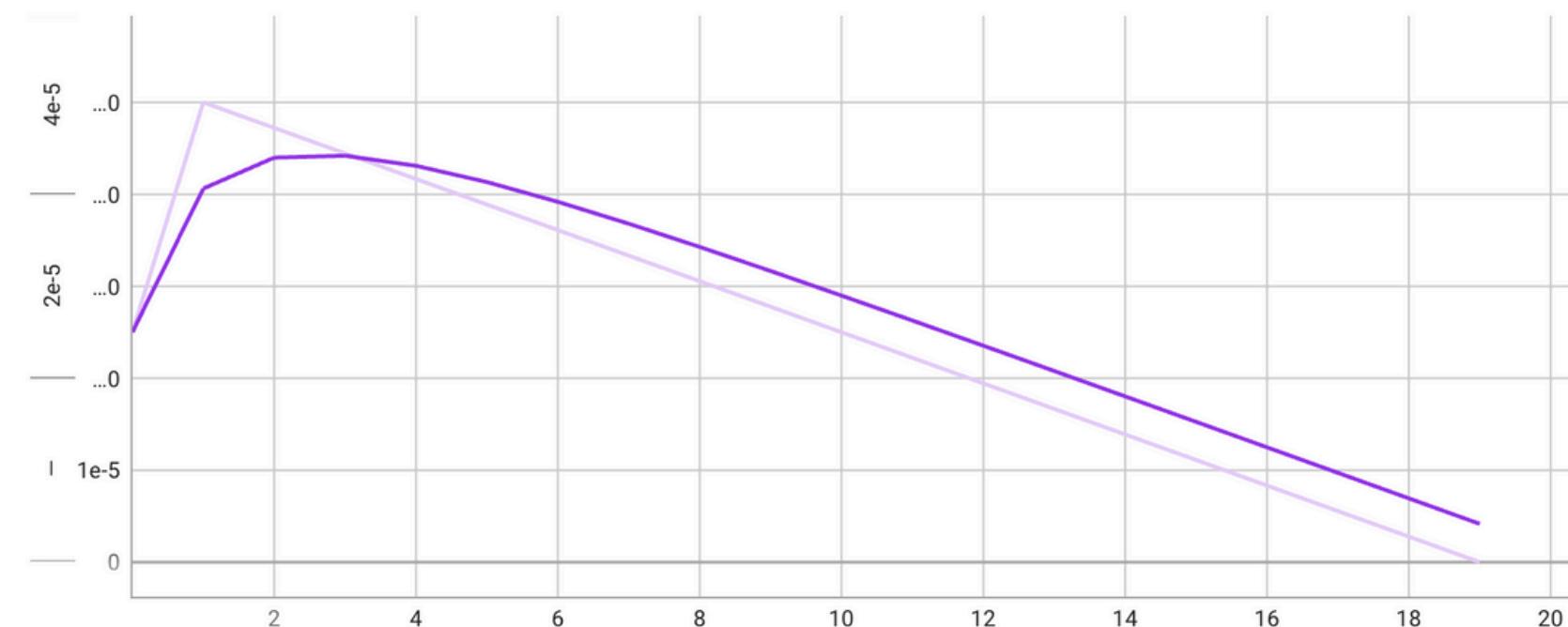
Esperimento 3



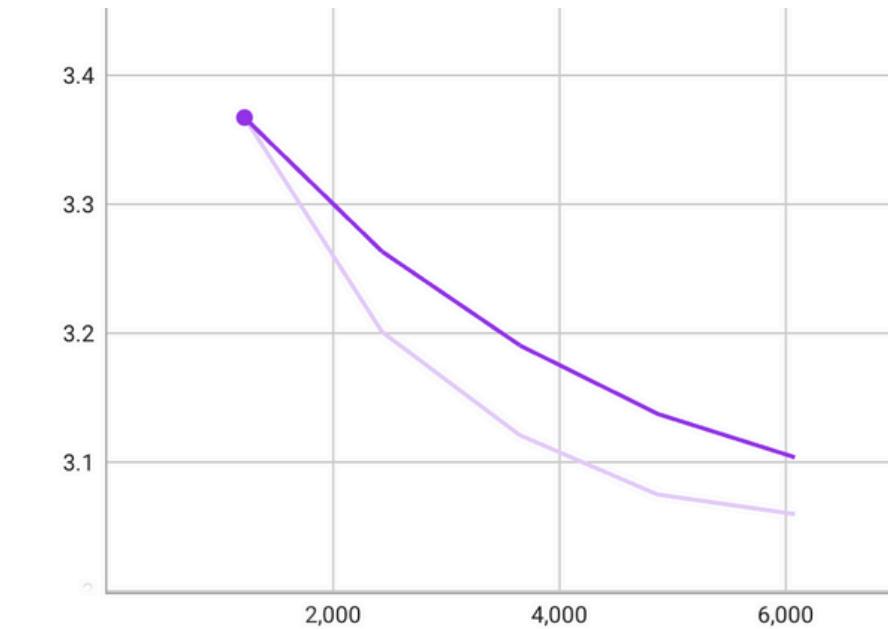
Categoria	Metrica / Componente	Precision	Recall	F1-Score
Generale	Validation Loss	3.0597		
NLG (Task RDF2Text)	BLEU	0.0357		
	METEOR	0.1502		
	ROUGE-L	0.1837		
RDF a Livello di Entità	Subjects	0.0000	0.0000	0.0000
	Predicates	1.0000	0.2118	0.3495
	Objects	0.0000	0.0000	0.0000
MLM	Accuracy (Soft)	0.0370		

Esperimento 3

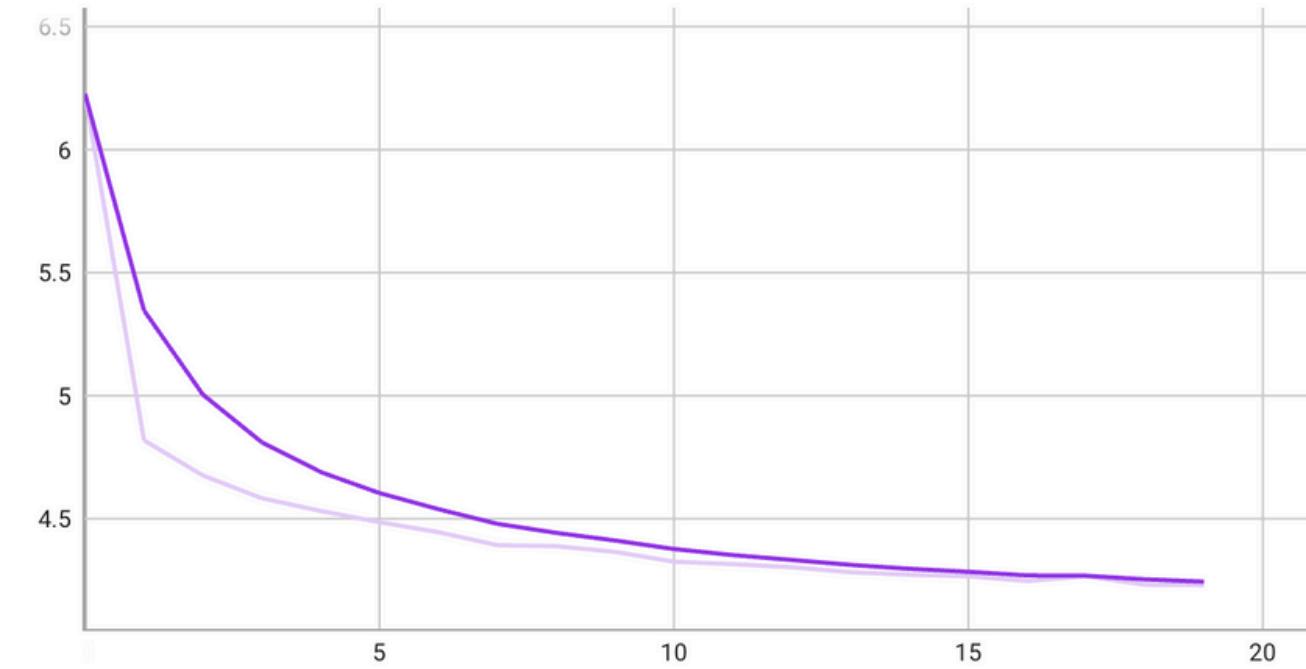
learning rate/decoder_tune



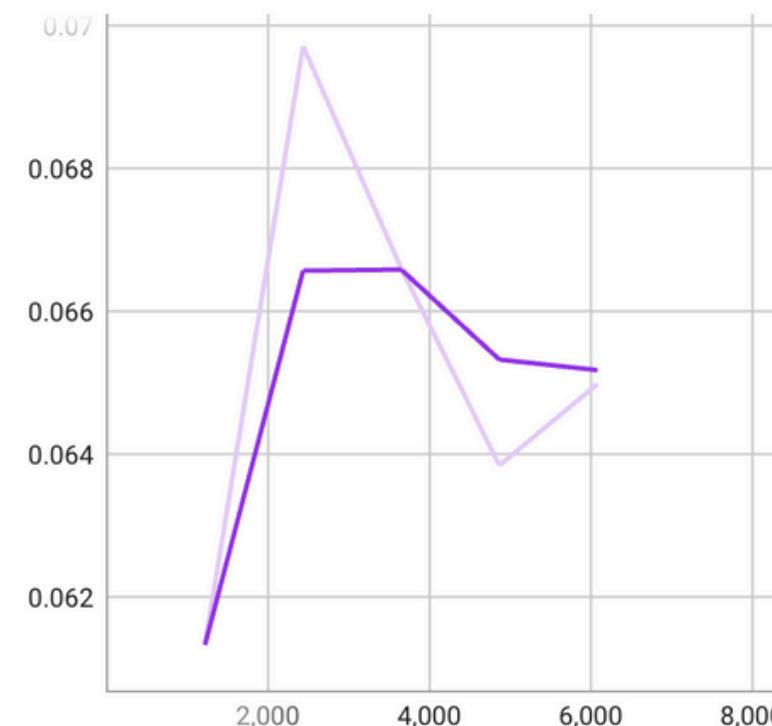
training_loss/decoder_tune



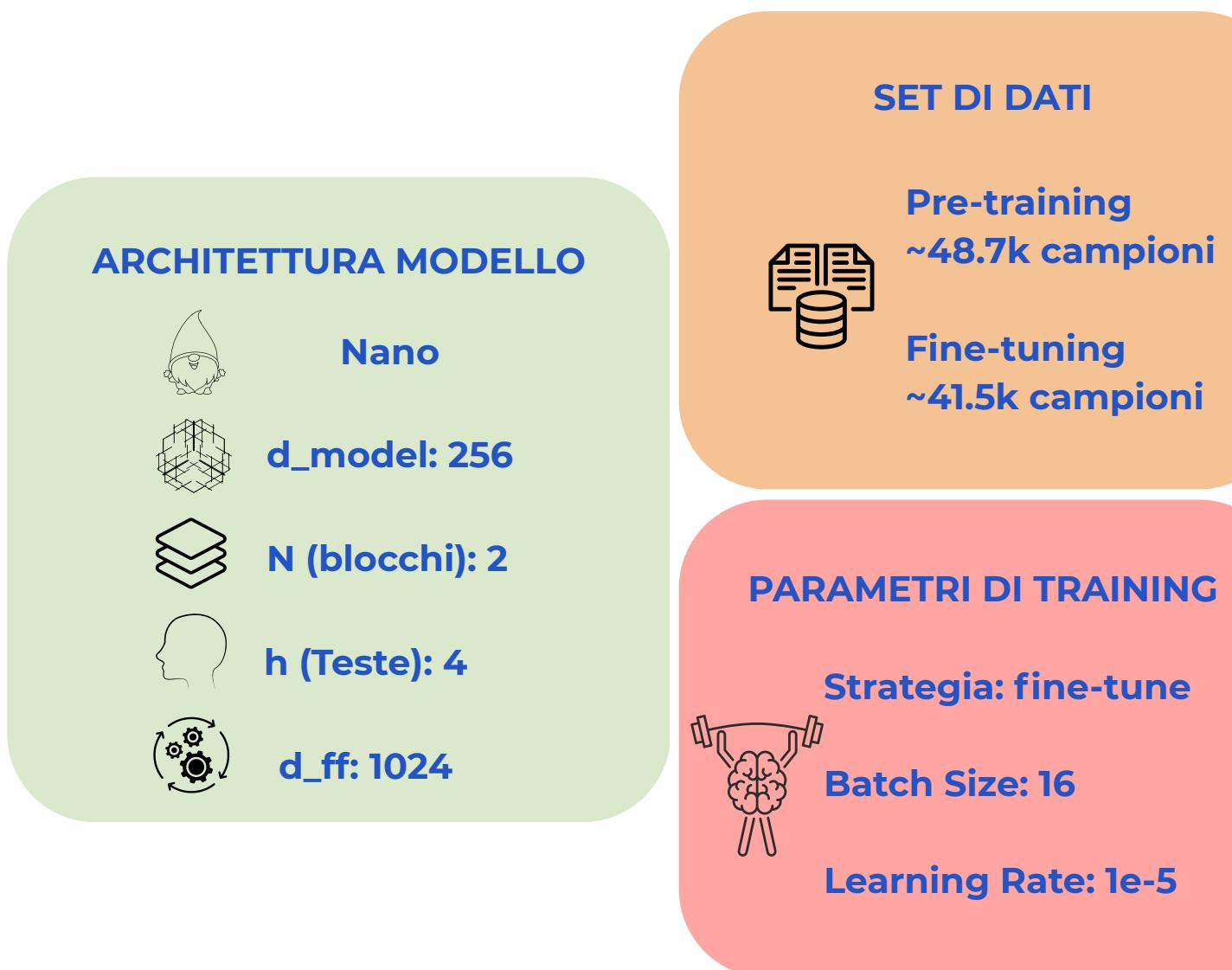
validation_loss/decoder_tune



token_f1/decoder_tune



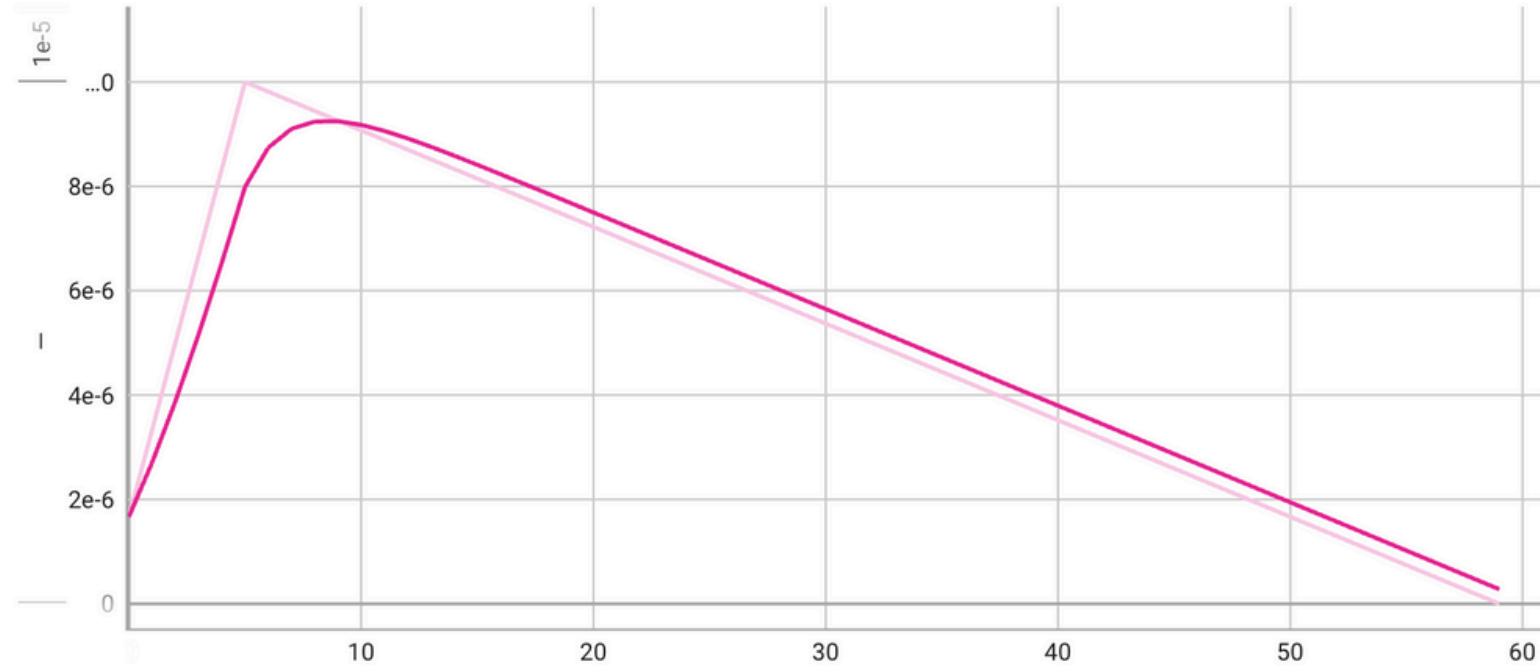
Esperimento 4



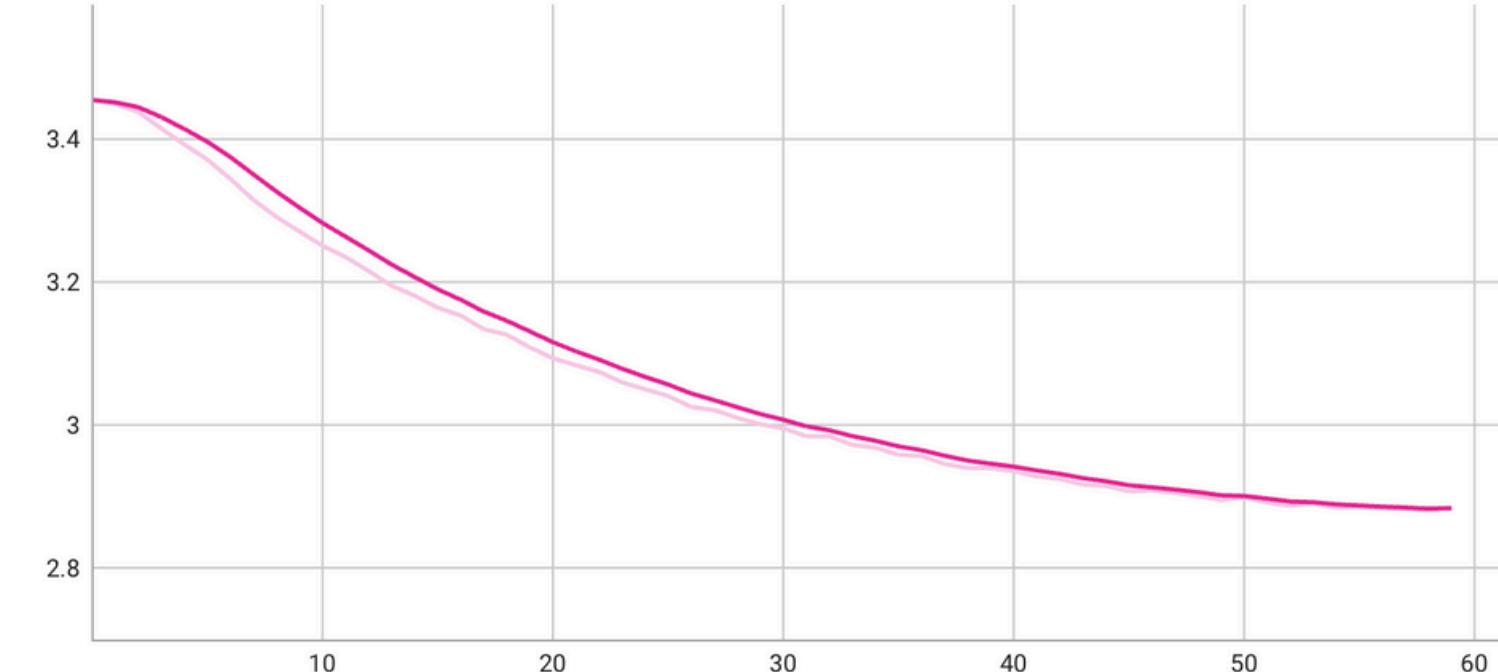
Categoria	Metrica / Componente	Precision	Recall	F1-Score
<i>Generale</i>	Validation Loss	1.7811		
<i>NLG (Task RDF2Text)</i>	BLEU	0.0419		
	METEOR	0.1948		
	ROUGE-L	0.2366		
<i>RDF a Livello di Entità</i>	Subjects	0.2738	0.0778	0.1212
	Predicates	0.9899	0.2813	0.4381
	Objects	0.2198	0.0625	0.0973
<i>MLM</i>	Accuracy (Soft)	0.2419		

Esperimento 4

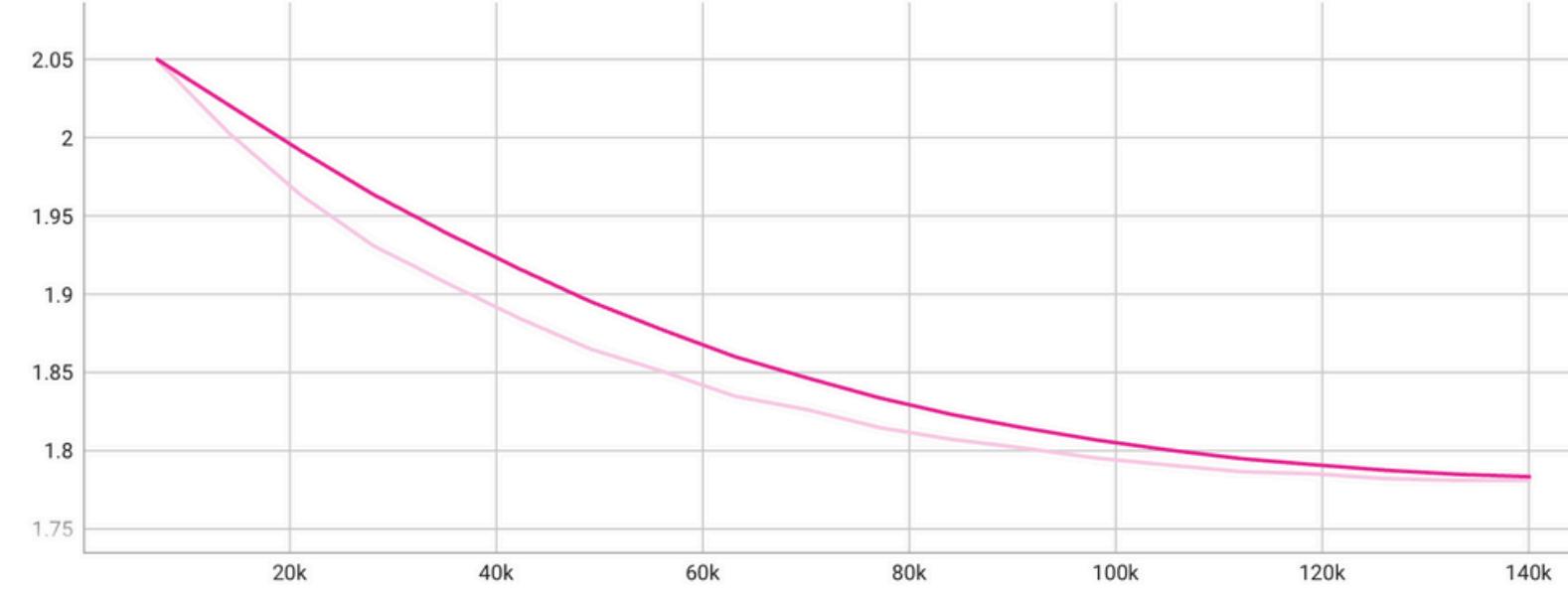
learning rate/full_finetune



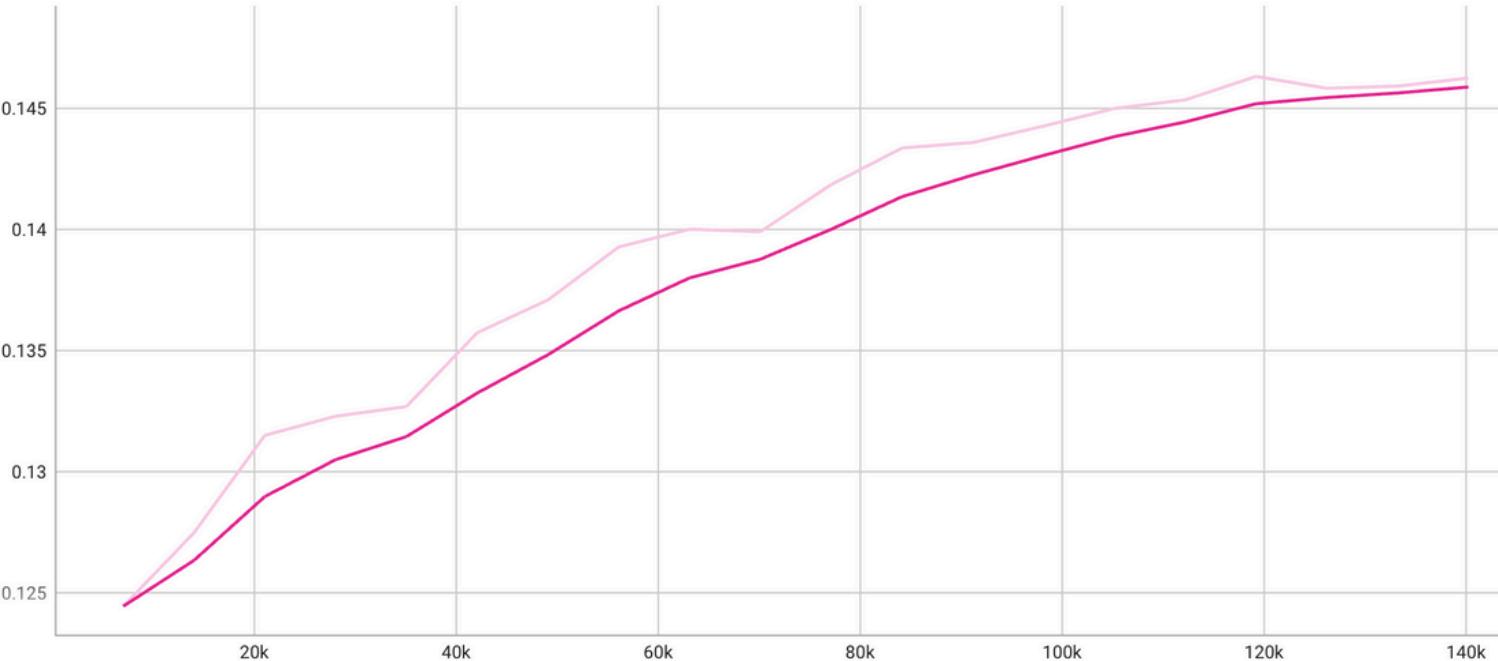
training_loss/full_finetune



validation_loss/full_finetune



token_f1/full_finetune



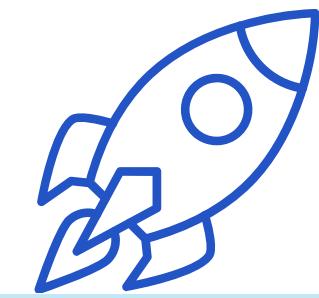
Conclusioni



L'equilibrio è tutto. Le performance del modello dipendono criticamente dalla sinergia tra capacità parametrica e scala dei dati. Ridurre uno dei due fattori porta a un inevitabile degrado.



Abbiamo costruito un modello semantico end-to-end compatto (stile T5) capace di operare sul dominio ibrido testo-RDF, dimostrando la fattibilità metodologica del nostro approccio.



Efficienza e Scalabilità. La nostra roadmap si concentra su tre pilastri per il futuro:

- PEFT (LoRA) per un fine-tuning ultra-efficiente.
- Stabilizzazione del Training con unfreezing granulare.
- TAPT per un pre-training più mirato.

**Cosa abbiamo imparato da
questo progetto?**



Grazie per l'attenzione!

