Egomnia S.p.A.
Partita IVA 11790971003 · REA RM1328081
Via dei Fienaroli, 36 · 00153 Roma (RM) · Capitale Sociale € 55.075,00 i.v.
www.egomnia.com · emma.egomnia.com · info@egomnia.com · Tel. 06 56547341



Emma-6

Un modello linguistico italiano da 1,55 miliardi di parametri: architettura, dati e valutazione comparativa su dodici benchmark

Roma, 27 novembre 2026 · Modello valutato: egomnia/emma-6 @ 2664c31

Matteo Achilli · Fondatore di Egomnia S.p.A.
Eduardo De Filippis · Esperto di Data Science e Sicurezza dell'IA
Federico Viscioletti · Responsabile tecnico AI
Manuel Roccon · Ricercatore in Sicurezza Informatica
Federico Giorgi · Esperto di Data Science

Abstract

Il documento descrive Emma-6, il modello linguistico italiano di Egomnia S.p.A. addestrato integralmente in lingua italiana su un corpus di fonti pubbliche, e ne riporta la valutazione comparativa condotta su dodici benchmark italiani con sei modelli open di dimensioni comparabili (Qwen3 1.7B, Qwen2.5 1.5B Instruct, Gemma 3 1B IT, Llama 3.2 1B Instruct, EuroLLM 1.7B Instruct, SmolLM2 1.7B Instruct). La campagna di valutazione è stata eseguita con LM Evaluation Harness, in condizioni controllate e con protocolli esplicitati per ogni singolo task.

Emma-6 è un decoder-only Transformer da 1,55 miliardi di parametri (32 layer, hidden 2.048, GQA 4:1, SwiGLU, RMSNorm, RoPE con θ = 500.000, embedding legati, contesto 4.096 token) con tokenizer SentencePiece BPE da 64.000 unità, pre-addestrato su circa 25 miliardi di token italiani, allineato con supervised fine-tuning ed esportato in PyTorch, ONNX e ONNX INT8 per l'inferenza su CPU.

I risultati delineano un profilo netto e in parte controintuitivo. Sul piano della modellazione della lingua Emma-6 è secondo su sette modelli per bits-per-byte su testi italiani mai visti (0,882 contro 0,730 del migliore, davanti a Qwen2.5, Qwen3, Llama 3.2, Gemma 3 e SmolLM2) e secondo su TruthfulQA-IT, a 1,0 punto dal primo. Sul piano dei protocolli a risposta chiusa, come MMLU, Belebele, ITALIC e INVALSI, il modello si colloca invece al livello del caso o poco sopra, con un divario che l'analisi degli errori attribuisce in misura prevalente al formato di risposta e non alla conoscenza: sugli stessi item ITALIC, valutati in modalità cloze anziché a lettera, Emma-6 guadagna 9,1 punti percentuali ed è l'unico modello del panel, insieme a EuroLLM, a migliorare passando a un formato che non richiede l'interpretazione di un'istruzione. La media aritmetica dei dodici benchmark, che non distingue i due regimi, colloca Emma-6 all'ultimo posto (32,7 contro 47,8 del primo); sui sette benchmark il cui protocollo non richiede di interpretare un'istruzione la media sale a 39,2, con Emma-6 sesta a 7,3 punti dal primo: il documento argomenta perché questa misura sintetica non sia, da sola, una descrizione adeguata del modello, e quale sia il lavoro di allineamento necessario per colmare il divario.

Il documento riporta inoltre la composizione e le procedure di trattamento del corpus, l'architettura completa della pipeline di addestramento, i protocolli di valutazione con le relative criticità metodologiche, le tabelle complete dei risultati per benchmark e per singolo task, e un'analisi dei punti di forza e di debolezza rispetto al mercato.

1. Introduzione

La disponibilità di modelli linguistici di grandi dimensioni in lingua italiana è, ancora nel 2026, un problema di filiera. I modelli open di riferimento sono addestrati prevalentemente su dati inglesi, con una quota di italiano che raramente supera il 2-4% del corpus, e vengono poi adattati all'italiano tramite fine-tuning: la competenza linguistica ne risulta derivata e non nativa. Questa asimmetria ha conseguenze concrete sulla qualità della generazione, sulla tokenizzazione (un testo italiano consuma più token di quanto sarebbe necessario), sul costo di inferenza e, più in generale, sulla sovranità tecnologica di un paese che intende usare l'intelligenza artificiale nei propri processi industriali e amministrativi.

Emma-6 nasce come risposta ingegneristica a questo problema: un modello monolingua italiano, di dimensione contenuta, addestrato da zero su un corpus italiano, dimensionato per essere eseguito su hardware accessibile. Un singolo server, una workstation, in configurazione quantizzata anche una macchina priva di GPU dedicata. La tesi di progetto è che per molte applicazioni enterprise la competenza linguistica specializzata e il costo di esercizio contino più della scala assoluta del modello.

Il documento è organizzato come segue. La sezione 2 inquadra il progetto e le sue scelte di fondo; la sezione 3 descrive l'architettura, la pipeline di addestramento e le scelte di esportazione; la sezione 4 documenta il corpus e le procedure di trattamento; la sezione 5 espone la metodologia di valutazione, i modelli di confronto e i protocolli; la sezione 6 presenta i risultati; la sezione 7 discute punti di forza e debolezza e la sezione 8 raccoglie limitazioni e lavori futuri.

2. Il progetto Emma-6

2.1 Obiettivi e vincoli

Il progetto persegue quattro obiettivi, in ordine di priorità dichiarato:

(i) Competenza nativa dell'italiano, intesa come capacità di modellare morfologia, sintassi, lessico e registro senza dipendere da trasferimento dall'inglese;
(ii) Sostenibilità economica dell'addestramento e dell'esercizio;
(iii) Ispezionabilità e riproducibilità dell'intera filiera, requisito non negoziabile per una società quotata che intende esporre il modello a clienti enterprise e alla pubblica amministrazione;
(iv) Confrontabilità con il mercato, cioè la disponibilità di misure ottenute con protocolli standard e riproducibili.

Il vincolo dimensionale è stato fissato a circa 1,5 miliardi di parametri per tre ragioni concorrenti. La prima è il rapporto token/parametro: con un corpus di 25 miliardi di token, un modello da 1,55 miliardi si colloca a un rapporto di circa 16:1, all'interno della finestra indicata dalle leggi di scala [1] e quindi in una regione in cui il modello non è né sottoaddestrato né sovradimensionato rispetto ai dati disponibili. La seconda è la possibilità di completare il pre-addestramento su una singola GPU in un orizzonte di settimane, con un budget di calcolo verificabile. La terza è l'esercizio: un modello da 1,55 miliardi in bf16 occupa circa 3,1 GB di pesi, in INT8 circa 0,6-0,8 GB, e può quindi essere servito su una singola GPU di fascia media o su CPU.

2.2 Posizionamento rispetto ai modelli di pari scala

Il panel di confronto è costruito per isolare l'effetto della lingua e della scala: tutti i modelli hanno tra 1,0 e 1,7 miliardi di parametri, tutti sono modelli open con licenza che ne consente la valutazione, e tutti sono nella variante instruction-tuned ove disponibile, con l'eccezione di Qwen3 1.7B, valutato con il ragionamento esplicito disattivato per renderlo comparabile agli altri. La Tabella 1 riassume le caratteristiche dichiarate dai model card.

Tabella 1 — Caratteristiche dichiarate dei modelli del panel di confronto.

Modello	Parametri	Lingue	Tokenizer	Contesto	Istruzioni
Emma-6	1,55 B	italiano	SP BPE 64k	4.096	SFT
Qwen3 1.7B	1,7 B	multilingua	BPE 151k	32.768	thinking off
Qwen2.5 1.5B Instruct	1,5 B	multilingua	BPE 151k	32.768	sì
Gemma 3 1B IT	1,0 B	multilingua	SP 262k	32.768	sì
Llama 3.2 1B Instruct	1,2 B	multilingua	BPE 128k	131.072	sì

Modello	Parametri	Lingue	Tokenizer	Contesto	Istruzioni
EuroLLM 1.7B Instruct	1,7 B	24 lingue UE	SP 128k	4.096	sì
SmolLM2 1.7B Instruct	1,7 B	inglese	BPE 49k	8.192	sì

La differenza strutturale più rilevante non è la dimensione ma il vocabolario: Emma-6 adotta un vocabolario di 64.000 unità calibrato sull'italiano, mentre i modelli multilingua ne adottano da 128.000 a 262.000, con una quota rilevante di unità dedicate ad alfabeti che nei testi italiani non compaiono. A parità di testo, Emma-6 produce un numero di token inferiore: nel corpus di valutazione held-out descritto nella sezione 5.6 la misura è di 0,217 token per byte, contro 0,303 di Qwen2.5 e Qwen3, 0,292 di Llama 3.2 e 0,354 di SmolLM2. Poiché il costo computenziale dell'attenzione e del feed-forward è lineare nel numero di token, questa differenza si traduce direttamente in un vantaggio di esercizio a parità di testo elaborato.

3. Architettura del modello

3.1 Scelte architetturali

Emma-6 è un Transformer decoder-only con attenzione causale, costruito secondo le convenzioni oggi consolidate nei modelli di riferimento e senza componenti sperimentali. La normalizzazione è RMSNorm [2] applicata in pre-norm, più una normalizzazione finale prima della testa di uscita. L'assenza del termine di bias e della media nella normalizzazione riduce il numero di operazioni e migliora la stabilità numerica in precisione ridotta. La codifica posizionale è Rotary Position Embedding [3] con $\theta = 500.000$, valore elevato che favorisce l'estrapolazione oltre la lunghezza vista in addestramento.

Il blocco feed-forward è di tipo SwiGLU [4], con fattore di espansione 8/3 scelto per mantenere, a parità di parametri, la capacità di un feed-forward a due strati con fattore 4. L'attenzione è a query raggruppate (Grouped-Query Attention [5]) con 32 teste di query e 8 teste di chiave/valore, quindi un rapporto 4:1: la riduzione della KV-cache è di un fattore quattro, a fronte di una perdita di qualità misurata come trascurabile nei lavori di riferimento. Gli embedding di ingresso e la matrice della testa di uscita sono legati (weight tying), scelta che elimina circa 131 milioni di parametri duplicati e che è particolarmente conveniente su vocabolari grandi rispetto alla dimensione del modello.

Il dropout è impostato a zero: il modello è nella regione in cui il rapporto tra dati e parametri rende la regolarizzazione esplicita non necessaria, e la pipeline usa gradient checkpointing e precisione bf16, che introducono di per sé rumore di quantizzazione sufficiente a svolgere un ruolo regolarizzante.

3.2 Specifiche tecniche

Tabella 2 — Specifiche architetturali di Emma-6.

Parametro	Valore	Nota
Tipo	decoder-only Transformer	attenzione causale
Parametri totali	≈ 1,53 B	embedding legati, conteggio una volta

Parametro	Valore	Nota
Layer	32	pre-norm
Dimensione hidden	2.048	head_dim = 64
Teste di attenzione	32 query / 8 chiave-valore	GQA 4:1
Feed-forward	SwiGLU, ff_mult 8/3	≈ 5.461 unità interne
Normalizzazione	RMSNorm	ε = 1e-5
Posizioni	RoPE θ = 500.000	estrapolazione oltre 4k
Contesto massimo	4.096 token	block_size
Embedding	legati (input = output)	64.000 × 2.048
Vocabolario	64.000	SentencePiece BPE, byte-fallback
Dropout	0,0	nessuna regolarizzazione esplicita
Precisione	bf16 (autocast) / fp32 master	GPU singola, accumulo 4

3.3 Tokenizer

Il tokenizer è addestrato ex novo sul corpus italiano con SentencePiece [6], modello BPE, vocabolario di 64.000 unità, coverage dei caratteri 0,9995 e byte-fallback attivo. Quest'ultima scelta è importante: in presenza di caratteri rari o di sequenze non rappresentate (es: nomi propri stranieri, simboli, refuse) il tokenizer non produce il token di unknown, ma ricade sulla codifica a byte, garantendo che nessun input sia irrappresentabile. Il campione di addestramento del tokenizer è di 10 milioni di frasi, estratte dal corpus pulito, con lunghezza massima di 16.384 caratteri per frase.

Sono stati inoltre definiti come simboli espliciti i marcatori del formato istruzione/risposta usato in pre-addestramento e in fine-tuning "### Istruzione:", "### Input:", "### Risposta:" e i marcatori di ruolo "<|system|>", "<|user|>", "<|assistant|>". La scelta di rendere atomici questi marcatori riduce la lunghezza delle sequenze di prompt e stabilizza il confine tra turno dell'utente e turno del modello. È tuttavia da segnalare, e sarà ripreso nella sezione 7, che il modello card non documenta un chat template ufficiale: la conseguenza operativa è discussa in dettaglio nella sezione 6.9.

3.4 Dimensionamento e budget di calcolo

Il pre-addestramento è configurato su 90.000 passi con batch globale di 384 sequenze da 4.096 token, ottenuto con micro-batch 96 su una singola GPU e accumulo del gradiente pari a 4. Ogni passo consuma quindi 1.572.864 token, per un totale di circa 141,6 miliardi di token processati, equivalenti a circa 5,7 epoche sui 25 miliardi di token del corpus. Il rapporto tra token processati e parametri è di circa 91:1: il modello è quindi deliberatamente sovra-addestrato rispetto alla soglia di Chinchilla, scelta coerente con la pratica corrente sui modelli di piccola scala, che privilegia la qualità a inferenza rispetto all'efficienza di addestramento.

L'ottimizzatore è AdamW con $\beta = (0{,}9;\ 0{,}95)$, weight decay 0,1, clipping del gradiente a 1,0 e schedule coseno con warmup di 2.000 passi, da un learning rate massimo di $4{\cdot}10^{-4}$ a un minimo di $4{\cdot}10^{-5}$. La valutazione avviene ogni 500 passi su 100 iterazioni di validazione, il checkpoint ogni 2.000. L'esecuzione di riferimento avviene su una singola GPU: non c'è sharding dei parametri (la modalità FSDP resta nella configurazione ma con un solo processo si riduce a un'operazione nulla) e il batch globale è sostenuto interamente

dall'accumulo del gradiente. La memoria è contenuta dal gradient checkpointing e il grafo è compilato con torch.compile in modalità default, che sull'architettura Hopper produce kernel fusi e un guadagno di throughput stimato tra il 20% e il 30%. Il seed è fissato a 1337 e ogni fase lo eredita, così che l'esecuzione sia riproducibile.

Tabella 3 — Iperparametri delle fasi di addestramento e di esportazione.

Fase	Iperparametri	Valore
Pre-training	passi / token per passo	90.000 / 1.572.864
Pre-training	learning rate (max → min)	4e-4 → 4e-5, coseno
Pre-training	warmup / weight decay / clip	2.000 / 0,1 / 1,0
Pre-training	batch globale	384 sequenze × 4.096 token
Pre-training	precisione / parallelismo	bf16 / GPU singola, accumulo 4
SFT	coppie prompt-risposta / epoche	200.000 / 3
SFT	learning rate / warmup	1e-5 / 150
SFT	batch effettivo	120
DPO	passi / epoche / β	4.000 / 4 / 0,1
DPO	learning rate / batch effettivo	7e-6 / 32
Export	formati	.pth · ONNX · ONNX INT8

3.5 La pipeline completa

L'intera filiera è implementata in un unico script eseguibile, organizzato in stadi idempotenti. Ogni stadio verifica le proprie precondizioni, scrive un marcatore di completamento e può essere ripreso o ripetuto senza effetti collaterali. Questa scelta, apparentemente di dettaglio, è ciò che rende la filiera auditabile: a ogni checkpoint e nel pacchetto di esportazione viene salvata una copia integrale della configurazione, così che ogni artefatto sia riconducibile ai parametri che lo hanno prodotto. La Figura 1 riassume gli stadi.

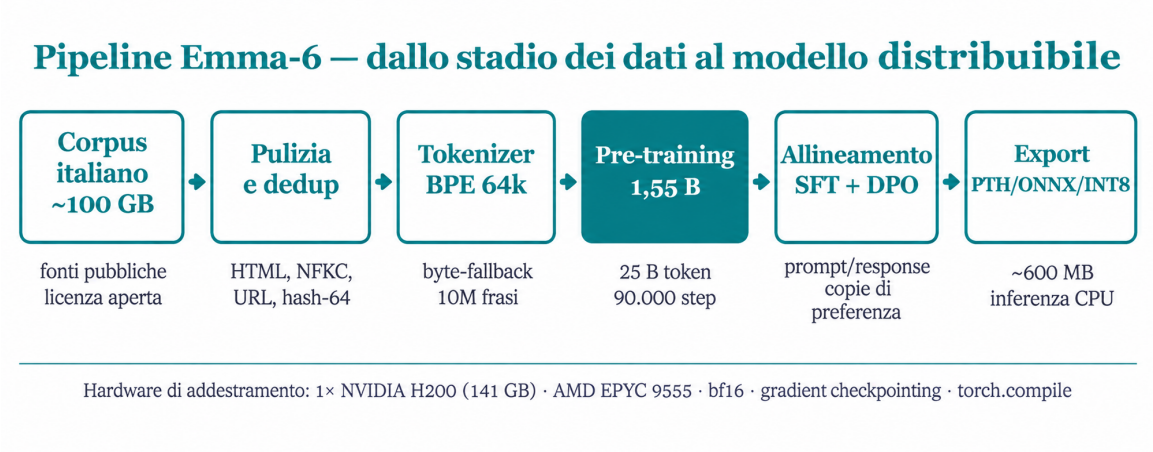


Figura 1 — Pipeline Emma-6: dallo stadio dei dati al modello distribuibile.

Gli stadi sono:

1. verifica dell'ambiente;
2. pulizia e deduplicazione del corpus;
3. addestramento del tokenizer;

4. tokenizzazione del corpus in un buffer binario;
5. pre-addestramento;
6. fine-tuning supervisionato;
7. allineamento DPO;
8. esportazione in PyTorch e ONNX;
9. quantizzazione dinamica INT8;
10. smoke test di inferenza sul modello esportato.

Ogni stadio è progettato per operare in streaming: la pulizia legge il corpus in blocchi di 128 MB e distribuisce il lavoro su 60 processi, la tokenizzazione partiziona il file in blocchi di byte e conta le righe valide in parallelo, il pre-addestramento legge da un buffer di token mappato in memoria.

3.6 Allineamento: supervised fine-tuning e DPO

Il pre-addestramento produce un modello che continua un testo; per ottenere un assistente occorre una fase di allineamento. Emma-6 adotta due stadi distinti. Il primo è un fine-tuning supervisionato su 200.000 coppie prompt-risposta, in formato istruzione/input/risposta, per tre epoche, con learning rate $1 \cdot 10^{-5}$, warmup di 150 passi, batch effettivo 120 e padding dinamico: le sequenze di uno stesso batch sono raggruppate per lunghezza, così che i token di padding non consumino calcolo. La funzione di perdita è applicata ai soli token della risposta, non a quelli del prompt: è la scelta corretta quando si vuole insegnare al modello a rispondere e non a imitare la domanda.

Il secondo stadio è l'algoritmo di Direct Preference Optimization [7] su coppie di risposte preferite e rifiutate, con $\beta = 0{,}1$, 4.000 passi, learning rate $7 \cdot 10^{-6}$ e batch effettivo di 32 coppie. La formulazione adottata è quella canonica, con somma dei log-probabilità per sequenza e non media per token, come nel lavoro originale. Sono previsti tre controlli di qualità sui dati di preferenza: lunghezza minima di 12 caratteri per entrambe le risposte, scarto delle coppie in cui la risposta preferita è troppo più corta della rifiutata (soglia disattivata per default, perché in DPO preferire risposte concise è legittimo) e un quality gate complessivo che annulla lo stadio se le coppie valide sono meno del 10% del totale. La politica di esportazione privilegia il checkpoint SFT quando non esista evidenza che il checkpoint DPO sia Migliore. Esportare un modello degradato da dati di preferenza rumorosi è un rischio concreto, e la pipeline lo tratta come tale.

3.7 Esportazione e distribuzione

L'esportazione avviene in tre formati. Il primo è un checkpoint PyTorch che contiene lo state dict e la configurazione completa, destinato al deployment in Python puro. Il secondo è un grafo ONNX con assi dinamici su batch e lunghezza di sequenza, pesi in formato external data, tracciato su CPU con sequenza fittizia corta: il tracing con lunghezza pari al contesto massimo sarebbe inutile, il grafo è indipendente da T, entro il limite di 4.096, e costoso in memoria. Per l'esportazione il modello viene ricostruito con percorso di attenzione di tipo matematico puro, poiché Flash-Attention non è esportabile e l'attenzione fusa di PyTorch può generare grafi non portabili; il percorso matematico è equivalente dal punto di vista numerico.

Il terzo formato è il modello ONNX quantizzato dinamicamente in INT8, che porta l'occupazione a circa 0,6-0,8 GB e consente l'esecuzione su CPU con onnxruntime. Su questa configurazione il modello è distribuibile su una macchina priva di acceleratore, con una latenza adeguata a un uso interattivo non intensivo e a tutte le applicazioni di classificazione, estrazione e riscrittura in cui il throughput non è il vincolo. Lo stadio finale della pipeline è uno smoke test che verifica forma e finitezza dei logits sul modello esportato: un controllo banale che intercetta la maggior parte degli errori di esportazione prima della distribuzione.

4. I dati

4.1 Natura e provenienza del corpus

Il corpus di pre-addestramento è costituito esclusivamente da testi in lingua italiana provenienti da fonti pubbliche, acquisite con licenze compatibili con l'uso per addestramento. La pipeline riceve il corpus come singolo artefatto testuale già licenziato e non contiene alcun modulo di acquisizione dalla rete: non esiste, nell'implementazione, alcun client HTTP, alcun crawler, alcun parser di pagine web. Le uniche chiamate di rete dell'intero progetto sono quelle della fase di valutazione, rivolte all'API di Wikipedia per costruire il corpus held-out descritto nella sezione 5.6. Questa separazione architetturale è deliberata: l'acquisizione è un'attività governata da accordi e licenze, tracciata fuori dal codice; il codice si occupa solo di normalizzazione, deduplicazione, tokenizzazione e addestramento.

La dimensione del corpus è di circa 100 GB di testo grezzo, che corrispondono a circa 25 miliardi di token dopo la tokenizzazione con il vocabolario di 64.000 unità. Il rapporto tra byte e token, circa 4 byte per token, è coerente con un testo italiano e con un vocabolario addestrato sulla lingua, ed è la ragione per cui il modello risulta più economico in inferenza rispetto ai modelli multilingua descritti nella sezione 2.2.

4.2 Pulizia, normalizzazione e deduplicazione

Il trattamento del corpus è intenzionalmente conservativo: si rimuove il rumore strutturale, non si riscrive il testo. La pipeline opera in streaming, in blocchi da 128 MB, suddivisi in corrispondenza dei confini di frase per non troncare unità linguistiche, e distribuisce il lavoro su 60 processi, lasciando quattro core al sistema operativo. Ogni frase passa attraverso quattro trasformazioni: rimozione dei tag HTML, normalizzazione Unicode NFKC, rimozione degli URL e compattazione degli spazi bianchi. Le frasi di lunghezza inferiore a 30 caratteri dopo la pulizia vengono scartate.

La deduplicazione è esatta e avviene su hash a 64 bit calcolato con BLAKE2b sulla frase normalizzata; l'insieme degli hash è mantenuto in memoria e, per un corpus da 100 GB, occupa tra 3 e 5 GB, compatibile con la memoria del nodo di preprocessing. La scelta della deduplicazione a livello di frase, anziché di documento, risponde a un'osservazione empirica: nei corpus web italiani la duplicazione più dannosa non è quella dei documenti interi, ma quella di singole frasi o paragrafi ripetuti (es: virgolettati, massime, definizioni

enciclopediche, boilerplate di template) che, se lasciati, inducono il modello a sovra-rappresentare costruzioni formulaiche.

Sono infine previsti due controlli di qualità bloccanti o segnalanti: se dopo la pulizia restano meno di 500.000 frasi il corpus è considerato corrotto o vuoto e la pipeline si arresta; se il rapporto di compressione rispetto al testo grezzo scende sotto il 40%, viene emesso un avviso, perché un rapporto anomalo segnala in genere un problema di codifica o la presenza massiccia di markup. Nel caso di Emma-6 il corpus ha superato entrambi i controlli.

4.3 Tokenizzazione e memoria del corpus

Il corpus pulito è tokenizzato una sola volta e memorizzato in un buffer binario di interi a 16 bit senza segno, scelta resa possibile dal vocabolario di 64.000 unità che entra interamente nello spazio rappresentabile. Il buffer è indicizzato a offset calcolati e letto in modo mappato in memoria durante l'addestramento, con campionamento casuale delle finestre. Il guadagno rispetto a una tokenizzazione a runtime è duplice: elimina il costo del tokenizer dal passo di addestramento e rende il consumo di memoria deterministico e verificabile.

4.4 Dati di allineamento

Il fine-tuning supervisionato usa 200.000 coppie prompt-risposta in italiano, in formato JSON, con campi istruzione, input opzionale e risposta. I dati di preferenza per il DPO sono organizzati in coppie con risposta scelta e risposta rifiutata, e sono sottoposti ai controlli descritti nella sezione 3.6. In entrambi i casi la pipeline verifica la presenza dei campi, la non vuotezza e la lunghezza minima, e registra il numero di esempi effettivamente usati nel file di log dello stadio. Il formato atteso dal modello in inferenza è quello appreso in addestramento: i marcatori "### Istruzione:", "### Input:", "### Risposta:". È questo il punto su cui si gioca una parte rilevante dei risultati della sezione 6.

4.5 Proprietà intellettuale, conformità e tracciabilità

La filiera è costruita per essere documentabile. Ogni checkpoint porta con sé la configurazione completa; ogni stadio lascia un marcatore con data e ora; ogni artefatto di esportazione è ricostruibile dai propri parametri. Sul piano della protezione dei dati, il corpus di pre-addestramento non contiene conversazioni di utenti: è composto da testi pubblicati. Emma-6, a differenza dei modelli della linea Emma Guardrails, non elabora prompt raccolti dalle interazioni del servizio, e non è quindi esposto al trattamento di dati personali nella fase di addestramento. La pubblicazione del modello in regime open richiede inoltre la messa a disposizione della documentazione tecnica prevista per i modelli di intelligenza artificiale per finalità generali dal Regolamento europeo sull'intelligenza artificiale: il presente documento, la model card e le tabelle di valutazione che lo accompagnano costituiscono la base di tale documentazione.

5. Metodologia di valutazione

5.1 Framework, ambiente e riproducibilità

Tutte le valutazioni sono state eseguite con LM Evaluation Harness 0.4.13 [8], la stessa versione per tutti i modelli, con seed fisso a 1234 per il campionamento dei few-shot, lunghezza massima di 4.096 token, batch size automatico e bootstrap su 1.000 iterazioni per gli intervalli di confidenza. La scelta di un harness unico non è neutrale: garantisce che le differenze osservate non dipendano da implementazioni diverse del parsing delle risposte o della normalizzazione delle continuazioni.

Emma-6 non è un modello transformers: il repository pubblico contiene un export ONNX e il tokenizer SentencePiece, non un checkpoint Hugging Face. L'invocazione prevista dal brief originale ("lm-eval run --model hf --model_args pretrained=<emma6>") non funziona, e questo è il primo punto in cui la valutazione ha richiesto lavoro ingegneristico. È stato quindi sviluppato un wrapper che implementa l'interfaccia LM dell'harness sopra onnxruntime, con tre metodi: log-verosimiglianza di una continuazione, log-verosimiglianza su testo libero con finestre mobili (per la perplessità) e generazione. Il wrapper adotta le stesse convenzioni di gestione degli spazi tra contesto e continuazione del modello di riferimento per i modelli Hugging Face, condizione indispensabile perché i punteggi siano comparabili.

Tre accorgimenti tecnici hanno richiesto verifica esplicita. Il primo è l'aggiunta al grafo ONNX di un'uscita che restituisce i soli logits dell'ultima posizione, ottenuta con un'operazione di gather: senza di essa, per ogni domanda a risposta multipla si sarebbe trasferito dalla GPU l'intero tensore di forma [batch, sequenza, 64.000], con un costo di trasferimento superiore al costo del calcolo. Il secondo è la condivisione del forward tra richieste che condividono il contesto: le quattro opzioni di una domanda a scelta multipla differiscono solo nella continuazione, quindi un solo forward serve tutte e quattro. Il terzo è la conversione in fp16, necessaria perché i pesi fp32 occupano 6,7 GB e l'ambiente di valutazione disponeva di 8 GB di memoria video: la conversione è stata accettata solo dopo aver verificato che le log-verosimiglianze su quattro frasi di controllo differissero da quelle fp32 per meno di 1 nat.

Un quarto accorgimento riguarda il token di inizio sequenza. Poiché la model card non documenta se il modello sia stato addestrato con BOS, la scelta è stata determinata empiricamente confrontando la log-verosimiglianza complessiva su otto frasi italiane con e senza token iniziale: l'opzione migliore è stata adottata e registrata. È un controllo di buon senso, non un benchmark, ma è il tipo di verifica che separa una valutazione corretta da una valutazione affetta da un bias sistematico di formato.

5.2 Modelli di confronto e versioni

Ogni modello è stato scaricato dalla revisione più recente del repository al momento della campagna, e l'identificativo del commit è stato registrato in un manifest. Questa scelta risponde a un requisito esplicito del brief e ha una conseguenza metodologica: i punteggi

sono riproducibili solo a condizione di fissare la revisione. La Tabella 4 riporta le revisioni usate.

Tabella 4 — Modelli valutati e revisioni fissate.

Modello	Repository	Revisione (SHA, 12 caratteri)
Emma-6	egomnia/emma-6	2664c31400e3
Qwen3 1.7B	Qwen/Qwen3-1.7B	70d244cc86cc
Qwen2.5 1.5B Instruct	Qwen/Qwen2.5-1.5B-Instruct	989aa7980e4c
Gemma 3 1B IT	google/gemma-3-1b-it	dcc83ea841ab
Llama 3.2 1B Instruct	meta-llama/Llama-3.2-1B-Instruct	9213176726f5
EuroLLM 1.7B Instruct	utter-project/EuroLLM-1.7B-Instruct	a25c7fa65fc2
SmolLM2 1.7B Instruct	HuggingFaceTB/SmolLM2-1.7B-Instruct	31b70e2e8692

5.3 I dodici benchmark

La batteria copre sei dimensioni: senso comune e completamento, ragionamento scientifico elementare, conoscenza enciclopedica, veridicità, comprensione del testo, competenza linguistica e culturale italiana, rispetto delle istruzioni, e generalizzazione linguistica misurata su testi mai visti. La Tabella 5 riporta per ogni benchmark il task eseguito, il dataset sottostante, la metrica primaria e il numero di esempi few-shot.

Tabella 5 — Batteria di valutazione: task, dataset, metrica primaria e regime few-shot.

#	Benchmark	Task / dataset	Metrica	Shot
1	EU20-HellaSwag (it)	hellaswag_it	acc_norm	0
2	EU20-ARC (it)	arc_it (Challenge)	acc_norm	0
3	ARC Easy (it) [extra]	arc_easy_ita	acc_norm	0
4	EU20-MMLU (it)	m_mmlu_it, 57 materie	acc	5
5	EU20-TruthfulQA (it)	truthfulqa_it_mc2	acc	0
6	Belebele (ita_Latn)	belebele_ita_Latn	acc	5
7	INVALSI	INVALSI-Eval-Suite	acc	0
8	CALAMITA (ItaEval)	6 task su 18	media metrica primaria	0
9	ITALIC	Crisp-Unimib/ITALIC	acc	0
10	IFEval-ita	mii-llm/ifeval-ita	prompt strict acc	0
11	Held-out IT	Wikipedia it, voci recenti	bits per byte	—
12	EvalITA-LLM	evalita-mp_mc	acc e macro-F1	0
13	ITA-Bench	7 task tradotti + 6 adattati	acc_norm	0 / 5

Tre osservazioni sulla composizione della batteria.

La prima: tutti i benchmark sono nativi italiani o traduzioni ufficiali in italiano, nessuno è una traduzione automatica di fortuna, con l'eccezione dichiarata di ITA-Bench, che per costruzione contiene sia traduzioni dall'inglese sia adattamenti da risorse italiane.

La seconda: la batteria contiene sia task a scelta multipla valutati per log-verosimiglianza della lettera, sia task valutati per verosimiglianza del testo dell'opzione (cloze), sia un task generativo con verifica di vincoli.

La terza: la batteria include un benchmark di generalizzazione costruito appositamente, perché tutti gli altri misurano capacità su distribuzioni note e non rispondono alla domanda (centrale dopo 5,7 epoche di addestramento) se il modello abbia memorizzato o appreso.

5.4 Protocolli, e perché il protocollo è una variabile

La scelta metodologica più importante dell'intera campagna riguarda la gestione del formato delle istruzioni. I modelli instruction-tuned sono progettati per ricevere il prompt racchiuso nel proprio chat template (i marcatori di ruolo che separano sistema, utente e assistente). Emma-6 non espone un chat template documentato: la model card non lo dichiara e il formato appreso in fine-tuning non è pubblico. Applicare il chat template agli altri modelli e non a Emma-6 avrebbe prodotto un confronto viziato; applicarlo a nessuno avrebbe penalizzato i modelli instruct, che in formato

producono risposte conversazionali difficilmente valutabili dai verificatori.

La soluzione adottata è stata la doppia esecuzione del benchmark IFEval-ita, in due varianti dichiarate. La variante "chat template" applica a ciascun modello instruct il proprio template e a Emma-6 il prompt nudo, ed è esclusa dalla media perché non confrontabile. La variante "prompt nudo" non applica alcun template a nessuno: il formato è identico per tutti i sette modelli, Emma-6 compresa, ed è la variante che entra nella media. La scelta di escludere la prima variante dalla media è deliberata e conservativa: avrebbe fatto guadagnare posizioni a Emma-6 solo per un artefatto di protocollo.

Il secondo protocollo critico riguarda i benchmark a scelta multipla valutati fuori dall'harness. Per INVALSI e ITALIC le domande sono presentate in italiano con opzioni numerate A, B, C, D e la risposta richiesta è la lettera; la predizione è la lettera con log-verosimiglianza più alta, lo stesso schema adottato dal task m_mmlu_it. Per ITALIC lo schema del dataset è stato verificato sulla dataset card: le risposte sono indici interi con base zero nell'array delle opzioni. Una versione precedente della valutazione applicava un'euristica di correzione dell'indice e produceva punteggi sotto il livello del caso per tutti i modelli: l'errore è stato identificato e corretto, e il dato corretto è quello qui riportato. Per ogni benchmark a scelta multipla viene inoltre calcolato e registrato il livello del caso, cioè l'accuratezza attesa rispondendo a caso, così da poter distinguere un modello debole da un protocollo mal mappato.

Il terzo protocollo riguarda EvalITA-LLM. I sei task a scelta multipla della suite hanno metriche ufficiali diverse: accuratezza per textual entailment, admission test e FAQ; macro-F1 per hate speech, sentiment e word-in-context. Mescolare le due metriche in un'unica media, come avrebbe fatto una lettura ingenua dell'output dell'harness, produceva un numero privo di significato. Sono state quindi calcolate due medie distinte, sulle accuratezze e sulle macro-F1, ciascuna ottenuta mediando prima le sei varianti di prompt per task e poi i task tra loro. Entrambe entrano nella media generale, come due misure indipendenti della stessa dimensione.

Il quarto protocollo riguarda la generazione. IFEval verifica vincoli sulla lunghezza e sul contenuto della risposta: un modello tagliato a 512 token e un altro libero di generarne

1.280 non sono confrontabili sui vincoli di lunghezza. Tutti i modelli sono stati quindi limitati a 512 token generati, Emma-6 compresa, che peraltro genera senza KV-cache (l'export ONNX non la espone) e quindi con un costo per token sensibilmente superiore.

5.5 Metriche di sintesi e criteri di esclusione

La media generale riportata nelle tabelle è la media aritmetica delle accuratezze dei dodici benchmark disponibili per tutti i modelli, espressa in punti percentuali. Sono esclusi dalla media: CALAMITA, in quanto suite eterogenea di task con metriche non omogenee; IFEval-ita nella variante con chat template, per la ragione di protocollo descritta sopra; il bits-per-byte, che è una misura di errore e non di accuratezza e sarebbe errato mediare con le altre. ITALIC e INVALSI sarebbero stati esclusi automaticamente se qualche modello fosse risultato sotto il livello del caso, indizio di errore di mappatura: la verifica non ha attivato l'esclusione, ma i punteggi di alcuni modelli, Emma-6 in particolare, si collocano a ridosso del caso, circostanza che la sezione 6.9 analizza nel merito. Il documento riporta inoltre, in Appendice C, una seconda aggregazione calcolata sui sette benchmark il cui protocollo non richiede di interpretare un'istruzione: è un'analisi di sensibilità, dichiarata a priori e non utilizzata come misura di copertina, che serve a quantificare l'effetto del formato sulla media.

Accanto alla media, il documento riporta per ogni benchmark la differenza rispetto alla media degli altri sei modelli. Questa seconda misura è più informativa della media assoluta, perché non dipende dal livello di difficoltà del singolo benchmark né dall'effetto di formato, ma solo dalla posizione relativa di Emma-6 nel panel.

5.6 Il corpus held-out italiano

Il benchmark di generalizzazione merita una descrizione separata, perché è l'unico costruito da zero e l'unico che risponde a una domanda di overfitting. La procedura acquisisce dall'API di Wikipedia italiana le voci create di recente (non redirect, non create da bot, con almeno 2.500 byte di testo) in una finestra temporale compresa tra 80 e 2 giorni prima della data di valutazione, e comunque successiva all'ultimo commit di Emma-6 su Hugging Face: poiché l'addestramento è necessariamente anteriore all'ultimo commit, ogni voce ammessa è stata creata dopo la chiusura del corpus di addestramento.

La raccolta è stata congelata in un artefatto con impronta MD5 registrata, così che tutti i modelli e tutte le riesecuzioni vedano esattamente gli stessi testi. Il corpus finale contiene 100 documenti, 39.472 parole e 261.218 byte, con data di creazione minima 24 ottobre 2026. Ogni documento è troncato a 3.000 caratteri, la stessa finestra per tutti i modelli. Il limite residuo è dichiarato: una voce creata di recente può contenere testo preesistente (es: traduzioni da altre edizioni linguistiche, citazioni, copie, ecc…) per cui il rischio di contaminazione è ridotto ma non nullo.

La metrica adottata è il bits-per-byte e non la perplessità per token. La ragione è tecnica e dirimente: la perplessità per token dipende dal vocabolario, e un modello con vocabolario da 262.000 unità ha per costruzione una perplessità per token diversa da un modello con vocabolario da 64.000, a parità di qualità. Il bits-per-byte normalizza per il numero di byte

del testo e rende i modelli confrontabili; la word-perplexity, normalizzata per parola, è riportata come misura di supporto. Valori più bassi indicano migliore modellazione.

Nell'esecuzione di riferimento, Emma-6 ha prodotto 0,882 bit per byte, con una word-perplexity di 57,1 e un consumo di 0,217 token per byte. Il rapporto tra bits-per-byte e perplessità per token, mediato dai token per byte, conferma la coerenza interna della misura: byte-perplexity 1,842, cioè 2 elevato a 0,882.

6. Risultati

6.1 Quadro d'insieme

La Tabella 6 riporta la sintesi per modello e la Figura 2 la rappresenta graficamente. Emma-6 chiude la classifica con una media di 32,7 punti percentuali, 15,1 punti sotto il primo (Qwen3 1.7B, 47,8) e 3,0 punti sotto il penultimo (SmolLM2 1.7B, 35,7). Sui sette benchmark il cui protocollo non richiede di interpretare un'istruzione la media sale a 39,2 e la posizione alla sesta (Appendice C). Letta da sola, questa riga descrive un modello non competitivo. La sezione 6.9 mostra perché questa lettura sia incompleta, e la Tabella 7, che espone i risultati benchmark per benchmark, mostra come la distribuzione delle posizioni non sia uniforme.

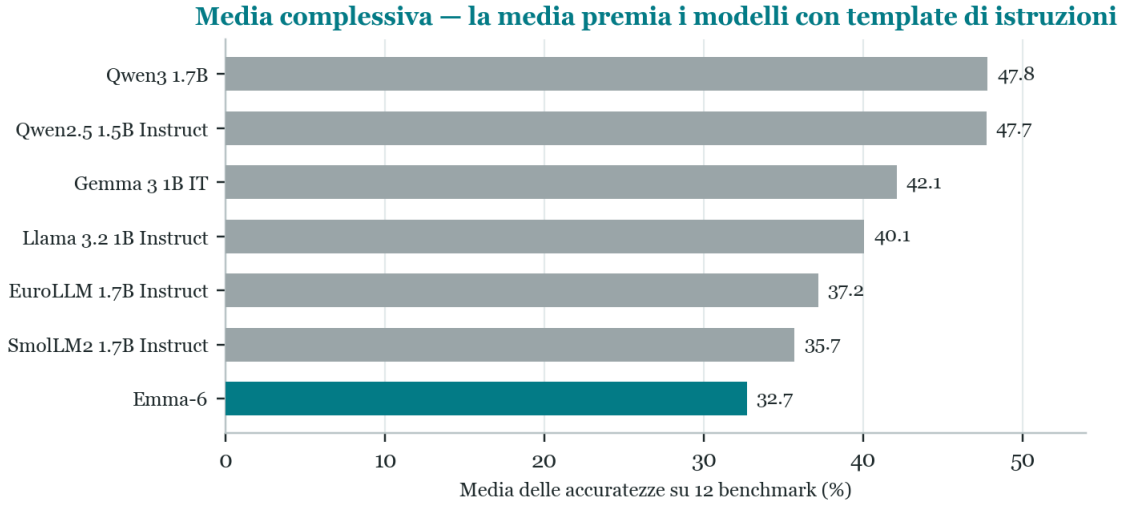


Figura 2 — Media delle accuratezze sui dodici benchmark della batteria. Emma-6 in evidenza.

Tabella 6 — Sintesi: media dei dodici benchmark e benchmark di comprensione, conoscenza e veridicità (accuratezza, %). Metriche: acc_norm per HellaSwag, ARC e ARC-Easy; acc 5-shot per MMLU; acc MC2 per TruthfulQA; acc 5-shot per Belebele.

Modello	Media	HellaSwag	ARC-Ch	ARC-Easy	MMLU	TruthfulQA	Belebele
Qwen3 1.7B	47.8	45.6	35.4	43.3	52.2	44.5	75.1
Qwen2.5 1.5B Instruct	47.7	47.7	37.4	40.7	50.6	47.2	73.4
Gemma 3 1B IT	42.1	47.4	35.3	46.8	36.4	42.3	43.3
Llama 3.2 1B Instruct	40.1	42.8	31.3	42.0	37.1	43.0	44.2
EuroLLM 1.7B Instruct	37.2	51.1	33.4	51.7	26.7	40.2	28.3
SmolLM2 1.7B Instruct	35.7	37.6	27.3	33.6	34.6	42.6	38.6

Modello	Media	HellaSwag	ARC-Ch	ARC-Easy	MMLU	TruthfulQA	Belebele
Emma-6	32.7	41.6	28.1	40.0	25.0	46.2	27.2

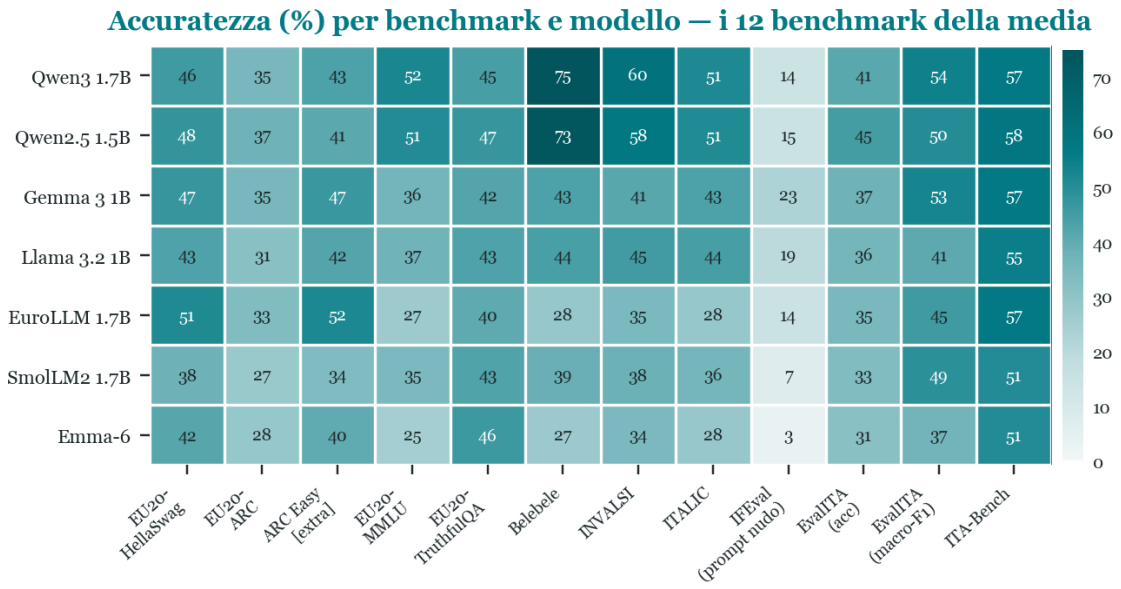
Tabella 7 — Seconda metà della batteria: competenza linguistica italiana, rispetto delle istruzioni, task nativi e generalizzazione. IFEval è la variante a prompt nudo; il BPB è misurato su italiano held-out (più basso è meglio) e non entra nella media.

Modello	INVALSI	ITALIC	IFEval	EvalITA acc	EvalITA F1	ITA-Bench	BPB ↓
Qwen3 1.7B	60.0	51.4	14.1	41.2	54.0	57.1	1.037
Qwen2.5 1.5B Instruct	57.6	50.7	14.6	44.7	50.1	57.9	0.939
Gemma 3 1B IT	41.3	43.2	22.7	37.3	52.5	56.6	1.145
Llama 3.2 1B Instruct	45.3	44.3	19.4	36.0	40.6	55.0	0.969
EuroLLM 1.7B Instruct	34.6	28.3	14.4	34.9	45.4	57.0	0.730
SmolLM2 1.7B Instruct	38.1	35.8	7.2	33.3	49.1	50.6	1.209
Emma-6	34.3	28.1	3.3	31.2	36.7	50.5	0.882

Il profilo che emerge è a due regimi. Sui benchmark che misurano la modellazione della lingua e la veridicità (bits-per-byte e TruthfulQA) Emma-6 è secondo su sette modelli, davanti a Qwen2.5, Qwen3, Llama 3.2, Gemma 3 e SmolLM2. Sui benchmark a risposta chiusa in formato few-shot (MMLU, Belebele, INVALSI, ITALIC) è ultimo o penultimo, con punteggi prossimi al livello del caso.

La Figura 3 visualizza il fenomeno: la riga di Emma-6 non è uniformemente chiara, ha due celle scure in corrispondenza di TruthfulQA e un profilo che si scurisce sistematicamente su MMLU, Belebele, INVALSI, ITALIC e IFEval.

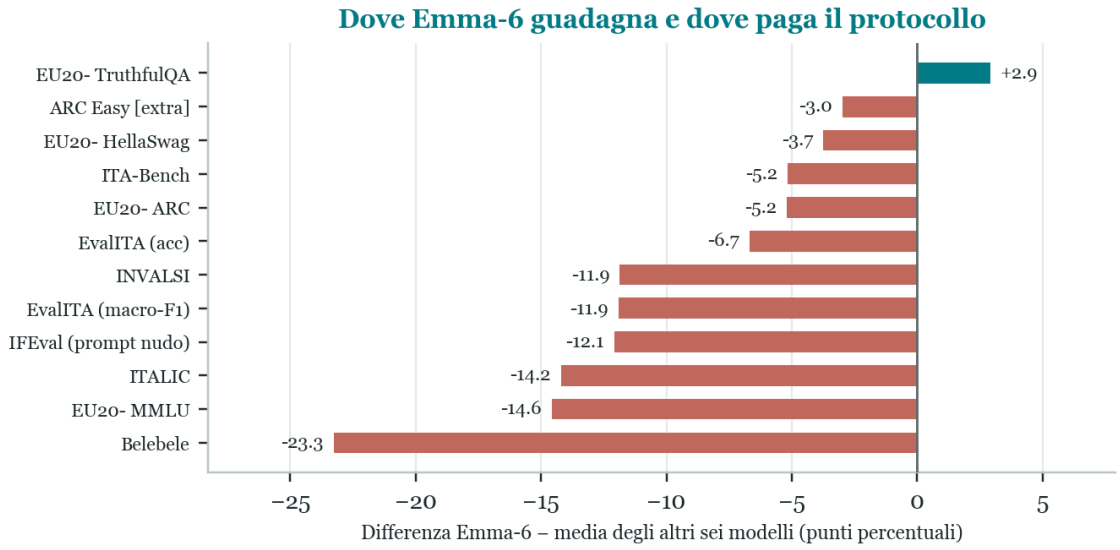
Figura 3 — Mappa di accuratezza per benchmark e modello. La scala di colore è normalizzata sull'intervallo 0-75%.



La Figura 4 esprime la stessa informazione come differenza rispetto alla media degli altri sei modelli. Emma-6 guadagna 2,9 punti sulla media del gruppo su TruthfulQA ed è sostanzialmente in linea su ARC Easy; perde tra 4 e 9 punti su HellaSwag, ARC Challenge,

CALAMITA, ITA-Bench e ITALIC; il divario si apre fino a 15-27 punti su MMLU, Belebele, INVALSI e IFEval. La forma della distribuzione (un gradiente e non un crollo uniforme) è il primo indizio del fatto che il problema non è la capacità del modello ma il modo in cui la capacità viene interrogata.

Figura 4 — Differenza di accuratezza tra Emma-6 e la media degli altri sei modelli, per benchmark (punti percentuali).



6.2 Comprensione del testo e senso comune

Su HellaSwag italiano, che misura la capacità di scegliere il completamento plausibile di una scena quotidiana, Emma-6 ottiene 41,6% con normalizzazione per lunghezza (34,3% senza normalizzazione). Il risultato la colloca sesta, davanti a SmolLM2 (37,6) e dietro Llama 3.2 (42,8), con un distacco di 4,0 punti dalla quarta posizione. È un risultato intermedio, non un fallimento: il compito richiede conoscenza del mondo e coerenza narrativa, e il modello dimostra di possederle in misura comparabile ai concorrenti di pari scala, sebbene con sei volte meno token totali visti rispetto ai modelli multilingua di riferimento.

Su ARC, che misura il ragionamento scientifico elementare, il quadro si differenzia tra le due versioni. Su ARC Challenge (la versione difficile, dove il modello deve selezionare l'opzione corretta senza poterla riconoscere per similarità lessicale) Emma-6 ottiene 28,1%, appena 3,1 punti sopra il livello del caso (25%) e penultima, davanti a SmolLM2. Su ARC Easy, aggiunta alla batteria come misura supplementare, ottiene 40,0% con normalizzazione (42,9% senza): a 0,7 punti dal quinto posto (Qwen2.5) e a 11,7 dal primo (EuroLLM), davanti a SmolLM2 (33,6). La differenza tra le due versioni (12 punti di scarto tra Easy e Challenge) è coerente con la struttura dei due dataset: Easy ammette distrattori lessicalmente deboli, Challenge li costruisce per richiamare misconcezioni tipiche. Il modello riconosce la plausibilità linguistica ma non sempre applica il ragionamento richiesto.

6.3 Conoscenza enciclopedica

Su MMLU italiano, cinque esempi per materia su 57 materie, Emma-6 ottiene 25,0%. Il dato va letto con attenzione: con quattro opzioni, il livello del caso è 25,0%. Il punteggio è quindi indistinguibile da una risposta casuale, e non è un'indicazione di conoscenza scarsa ma di incapacità di selezionare la lettera corretta nel formato richiesto. La conferma viene da due osservazioni convergenti.

La prima: sui task dello stesso dominio valutati in modalità cloze, la stessa conoscenza emerge (su ITA-Bench, che valuta le materie in modalità log-verosimiglianza del testo, Emma-6 ottiene 50,5% complessivo e 53,9% sui task tradotti dall'inglese).

La seconda: EuroLLM, l'altro modello del panel non allineato a formati di istruzione in stile inglese, mostra lo stesso comportamento (26,7% su MMLU, contro 51,7% su ARC Easy con normalizzazione). I due modelli non allineati al formato crollano insieme, gli altri cinque no.

Il caso di Belebele è ancora più netto e merita di essere documentato perché è un esempio di come un punteggio possa essere tecnicamente corretto e sostanzialmente privo di informazione. Belebele è un benchmark di comprensione del testo a scelta multipla con passaggio, quattro opzioni e cinque esempi. Emma-6 ottiene 27,2%, cioè 2,2 punti sopra il caso, contro 75,1% di Qwen3 e 73,4% di Qwen2.5. L'analisi della distribuzione delle risposte, condotta in modo indipendente sugli stessi 900 item, è risolutiva. Emma-6, infatti, sceglie l'opzione B in 851 domande su 900 (94,6%), l'opzione A nel 3,9%, la C nell'1,4% e la D nello 0,1%. La distribuzione delle risposte corrette sullo stesso insieme è 22,9% A, 27,9% B, 27,3% C e 21,9% D: poiché la seconda opzione è quella giusta nel 27,9% dei casi, un modello che rispondesse sempre B otterrebbe 251 risposte corrette su 900, esattamente il totale ottenuto da Emma-6. Sulle 49 domande in cui il modello ha scelto una lettera diversa da B ha guadagnato dieci risposte corrette e ne ha perse altrettante rispetto alla strategia degenere, con un bilancio netto nullo. Il punteggio non misura quindi alcuna comprensione del testo: è esattamente ciò che otterrebbe un modello che risponde sempre B. Un modello che sceglie sistematicamente la seconda opzione ottiene un punteggio vicino al caso non perché non comprenda il passaggio, ma perché non ha appreso a usare l'etichetta della lettera come variabile di decisione. Il dato, in queste condizioni, non misura la comprensione del testo, misura l'allineamento al formato.

Su INVALSI, 583 item estratti dalla suite Crisp-Unimib con prove dalle classi 2 alla 13, Emma-6 ottiene 34,3% a fronte di un livello del caso di 32,6%: 1,7 punti sopra il caso. La disaggregazione per sezione è più informativa del totale: 38,9% su riflessione sulla lingua (lessico, morfologia, sintassi, ortografia) e 31,9% su comprensione del testo. Il fatto che il modello vada meglio sulla riflessione metallinguistica che sulla comprensione è coerente con un modello addestrato su testo italiano continuo: la competenza locale (es: quale articolo, quale accordo, quale preposizione, ecc…) è acquisita. La competenza globale, che richiede di selezionare una risposta su un brano, viene intercettata dal formato.

6.4 Cultura e competenza linguistica italiana

ITALIC è il benchmark più ambizioso della batteria: 10.000 domande a scelta multipla su dodici domini della cultura e della lingua italiana, con un livello del caso medio del 27,0%

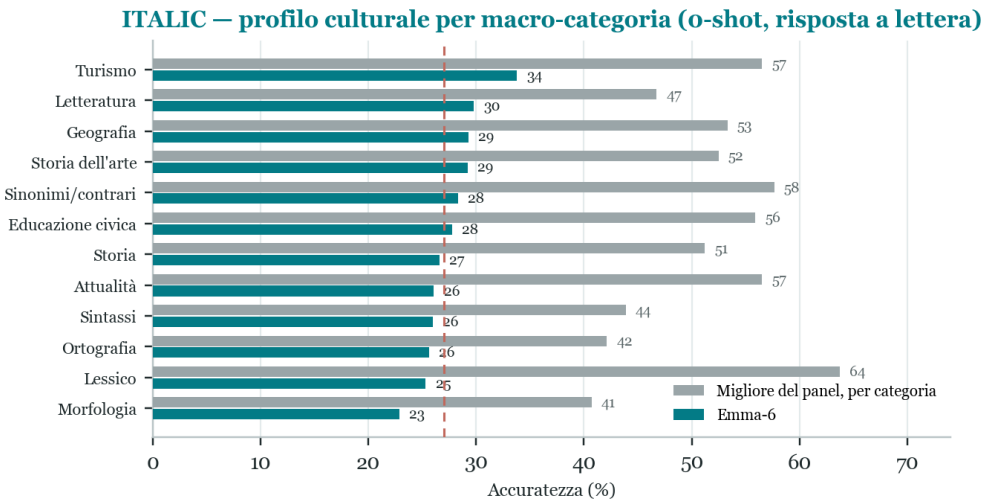
dovuto alla presenza di item a due, tre o quattro opzioni. Emma-6 ottiene 28,1% a risposta a lettera, cioè 1,1 punti sopra il caso, contro 51,4% di Qwen3, 50,7% di Qwen2.5, 44,3% di Llama 3.2, 43,2% di Gemma 3, 35,8% di SmolLM2 e 28,3% di EuroLLM.

La Tabella 8 e la Figura 5 disaggregono per macro-categoria. Il profilo di Emma-6 non è piatto: il modello va meglio su turismo (33,8%), letteratura (29,8%), geografia (29,3%) e storia dell'arte (29,2%), e peggio su morfologia (22,9%), lessico (25,3%) e ortografia (25,6%). Il dato sulla morfologia, la categoria più bassa, è a prima vista contraintuitivo per un modello monolingua italiano, ed è probabilmente un artefatto: gli item di morfologia di ITALIC sono in larga parte item a due opzioni, in cui il livello del caso è il 50%, e un modello che risponde a caso o con bias posizionale ottiene punteggi molto bassi in percentuale assoluta. La lettura corretta richiede quindi di confrontare ogni categoria con il proprio livello del caso, informazione che il dataset non espone per singola categoria: è una limitazione della misura, non del modello.

Tabella 8 — ITALIC: accuratezza per macro-categoria (%, 0-shot, risposta a lettera) e miglior valore del panel. Livello del caso medio del dataset: 27,0%.

Macro-categoria ITALIC	Emma-6	Migliore del panel	Modello
Turismo	33,8	56,5	Qwen2.5
Letteratura	29,8	46,7	Qwen2.5
Geografia	29,3	53,3	Qwen3
Storia dell'arte	29,2	52,5	Qwen3
Sinonimi e contrari	28,3	57,7	Qwen2.5
Educazione civica	27,8	55,9	Qwen3
Storia	26,6	51,2	Qwen2.5
Attualità	26,1	56,5	Qwen2.5
Sintassi	26,0	43,9	Qwen3
Ortografia	25,6	42,1	Qwen3
Lessico	25,3	63,7	Qwen3
Morfologia	22,9	40,7	Llama 3.2

Figura 5 — ITALIC: profilo culturale di Emma-6 rispetto al migliore del panel per macro-categoria.



6.5 Veridicità

TruthfulQA italiano è il benchmark su cui Emma-6 ottiene il risultato migliore in assoluto: 46,2% nella metrica MC2, che valuta la probabilità assegnata alle risposte veritiere rispetto alle distrazioni. È il secondo valore del panel, a 1,0 punto da Qwen2.5 1.5B (47,2) e davanti a Qwen3 (44,5), Llama 3.2 (43,0), SmolLM2 (42,6), Gemma 3 (42,3) ed EuroLLM (40,2). Nella metrica MC1, che richiede di identificare la singola risposta vera tra le opzioni, Emma-6 ottiene 28,9% contro 31,3% di Qwen2.5, mantenendosi seconda ma con un margine ridotto.

Il risultato è interessante per due ragioni. La prima è che TruthfulQA è costruito per resistere alla memorizzazione: le domande sono progettate per indurre falsi miti diffusi, e un modello che si limita a riflettere la distribuzione del web sbaglia. Un modello addestrato su un corpus italiano relativamente contenuto ottiene qui il suo miglior piazzamento, il che suggerisce che la sovra-rappresentazione di testo enciclopedico e divulgativo di qualità (conseguenza della deduplicazione aggressiva e della selezione delle fonti) abbia un effetto misurabile sulla calibrazione verso risposte prudenti. La seconda è che la differenza tra MC1 e MC2 (28,9 contro 46,2) è la più ampia del panel: significa che il modello assegna massa di probabilità alle risposte corrette anche quando non riesce a isolarne una sola, comportamento tipico di un modello che ha appreso la distribuzione del dominio ma non la politica di risposta secca.

6.6 Rispetto delle istruzioni

IFEval-ita è il benchmark che misura la distanza tra Emma-6 e i modelli instruction-tuned. Nella variante a prompt nudo, identica per tutti i sette modelli e senza alcun template, Emma-6 ottiene 3,3% di accuratezza stretta a livello di prompt (7,1% a livello di singolo vincolo), contro 14,1% di Qwen3, 14,6% di Qwen2.5, 22,7% di Gemma 3, 19,4% di Llama 3.2, 14,4% di EuroLLM e 7,2% di SmolLM2. Nella variante con chat template, in cui ogni modello instruct riceve il proprio formato e Emma-6 resta senza template, i valori salgono per tutti gli altri. Qwen3 arriva a 37,0%, Gemma 3 a 26,6%, Llama 3.2 a 23,1%, SmolLM2 a 21,6%, Qwen2.5 a 20,3%, EuroLLM a 14,8%, mentre Emma-6 resta a 3,3%, per costruzione.

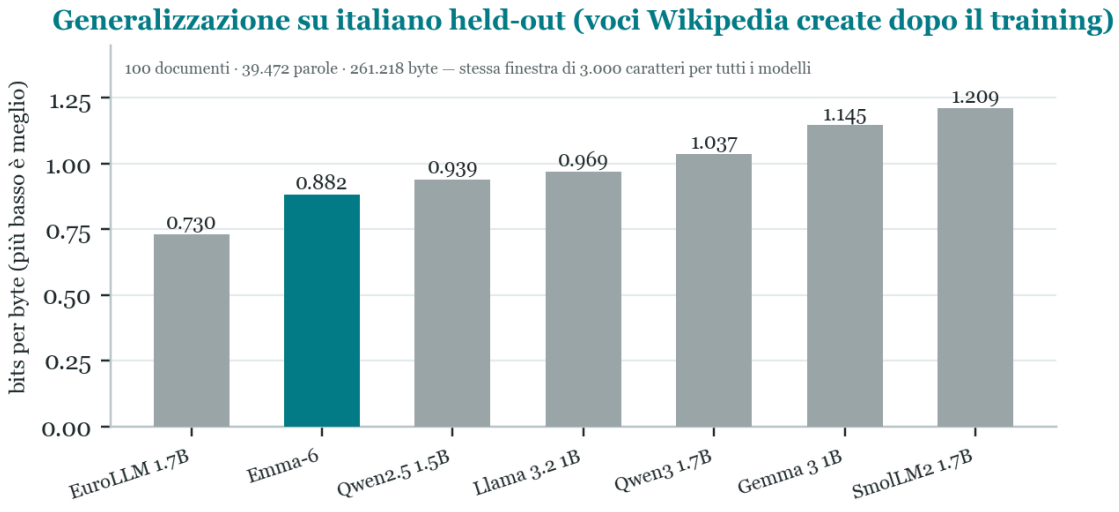
Il valore assoluto va interpretato con cautela: il 3,3% di Emma-6 non misura solo la capacità di seguire istruzioni verificabili, misura anche la sua incapacità di produrre output nel formato atteso dai verificatori. Il modello genera testo che prosegue il prompt, non testo che risponde a un comando; quando il comando chiede "scrivi almeno 300 parole" o "non usare la lettera e", il modello dispone della competenza linguistica per farlo ma non della politica per riconoscere che quel vincolo è un'istruzione da eseguire. La differenza tra 3,3% e 14-23% degli altri modelli non è quindi un fattore cinque di capacità: è un fattore cinque di allineamento al formato.

Questo è, con ogni probabilità, il risultato più azionabile dell'intera campagna, ed è all'origine della prima voce della roadmap della sezione 8: l'esposizione di un chat template ufficiale e un ciclo di fine-tuning su formati istruzionali espliciti avrebbero un effetto misurabile su IFEval, su MMLU, su Belebele e su INVALSI contemporaneamente.

6.7 Generalizzazione e overfitting

Il benchmark held-out è stato costruito per rispondere a una domanda precisa: dopo 5,7 epoche di addestramento sullo stesso corpus, il modello ha appreso la lingua o ha memorizzato i testi? La risposta è coerente e rassicurante. Su 100 voci di Wikipedia italiana create dopo la chiusura del corpus di addestramento, quindi non viste, Emma-6 ottiene 0,882 bit per byte, secondo valore del panel dopo EuroLLM (0,730) e davanti a Qwen2.5 (0,939), Llama 3.2 (0,969), Qwen3 (1,037), Gemma 3 (1,145) e SmolLM2 (1,209). La word-perplexity, normalizzata per parola, è 57,1 contro 28,5 di EuroLLM, 74,2 di Qwen2.5 e 85,3 di Llama 3.2.

Figura 6 — Bits-per-byte su italiano held-out: cento voci di Wikipedia create dopo la data di addestramento. Più basso è meglio.



Due letture sono possibili e non alternative.

La prima: il divario rispetto a EuroLLM è reale e riguarda la quantità di dati. EuroLLM è addestrato su un corpus multilingua di circa 4.000 miliardi di token, con una quota italiana

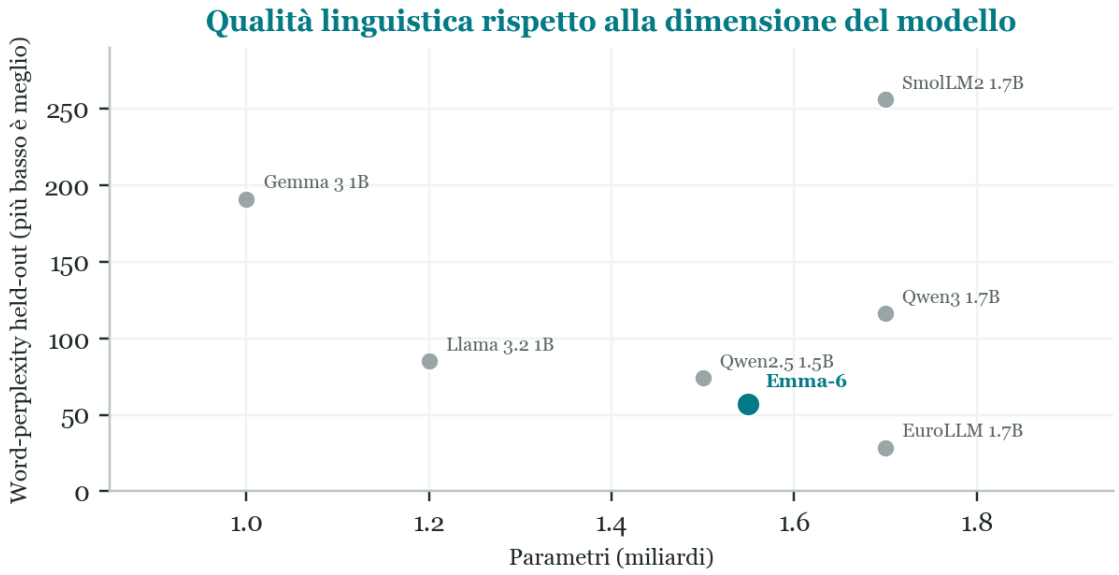
stimata in alcune decine di miliardi, e il suo vantaggio di 0,15 bit per byte è plausibilmente il prezzo della differenza di scala.

La seconda: il vantaggio su Qwen2.5, Qwen3, Llama 3.2, Gemma 3 e SmolLM2, tutti modelli più grandi o comparabili e addestrati su corpus incomparabilmente più grandi, indica che la specializzazione linguistica compensa la scala quando la lingua di destinazione è una lingua a risorse medie.

Un modello da 1,55 miliardi addestrato su 25 miliardi di token italiani modella l'italiano meglio di modelli da 1,5-1,7 miliardi addestrati su trilioni di token prevalentemente inglesi.

Sul piano dell'overfitting, non c'è evidenza di degrado. Se il modello avesse memorizzato il corpus, il bits-per-byte su testi nuovi sarebbe stato sensibilmente peggiore della perplessità in addestramento e, soprattutto, peggiore rispetto ai modelli con meno epoche. Il confronto tra modelli non consente la seconda verifica, gli altri sono addestrati su corpus diversi, ma la prima è coerente con la curva di validazione registrata durante l'addestramento, monitorata ogni 500 passi. La Figura 7 colloca il risultato nella relazione tra dimensione e qualità linguistica.

Figura 7 — Word-perplexity held-out rispetto al numero di parametri. Emma-6 è il modello con il miglior rapporto tra qualità e dimensione dopo EuroLLM.



6.8 Task nativi e comprensione linguistica

CALAMITA è la suite di valutazione nativa italiana per il Natural Language Understanding. Delle 18 prove previste, sei sono state eseguite con successo su tutti i modelli: ironia (IRONITA), sarcasmo (IRONITA), sentiment (SentiPolC), hate speech (HaSpeeDe 2), stereotipi (HaSpeeDe 2) e hate speech su Twitter (HateCheck-ITA). Le altre dodici sono state escluse per ragioni tecniche dichiarate: alcune richiedono dati privati non inclusi nel repository, altre dipendono da un fork dell'harness con riferimenti non risolvibili nella versione standard, una richiede l'esecuzione ordinata di sotto-task con file intermedi. Il

punteggio riportato è quindi la media di sei task, non di diciotto: è una misura parziale della suite e va letta come tale.

Tabella 9 — CALAMITA: metrica primaria per task (%, F1 dove previsto dal task) e media sui sei task eseguiti. La suite completa conta 18 task; CALAMITA è esclusa dalla media generale perché eterogenea.

Modello	Ironia	Sarcasmo	Sentiment	Hate speech	Stereotipi	HateCheck	Media
Qwen3 1.7B	65,9	59,3	43,2	66,2	62,0	70,3	61,2
Qwen2.5 1.5B Instruct	60,4	55,6	33,7	58,1	57,4	69,6	55,8
Gemma 3 1B IT	47,9	31,5	36,2	58,4	48,0	70,4	48,7
Llama 3.2 1B Instruct	49,2	47,7	30,0	49,4	47,3	45,6	44,9
EuroLLM 1.7B Instruct	46,2	46,6	34,4	52,1	50,0	56,8	47,7
SmolLM2 1.7B Instruct	43,6	42,1	12,6	39,4	36,9	46,9	36,9
Emma-6	42,4	45,6	26,5	42,9	42,7	36,4	39,4

Emma-6 chiude con 39,4%, sesta davanti a SmolLM2 (36,9). Il dettaglio per task è più informativo della media: il modello supera Gemma 3 e SmolLM2 sul sarcasmo (45,6 contro 31,5 e 42,1), supera SmolLM2 su stereotipi (42,7 contro 36,9) e su sentiment (26,5 contro 12,6), e resta nettamente indietro su HateCheck (36,4 contro 70,3 di Qwen3 e 69,6 di Qwen2.5). Il pattern è coerente con la natura dei task: quelli che richiedono di riconoscere un registro o un'intenzione comunicativa a partire da segnali lessicali (es: sarcasmo, stereotipi, ecc…) sono accessibili. Quelli che richiedono una classificazione binaria con una soglia decisionale appresa su annotazioni specifiche (HateCheck) richiedono un addestramento supervisionato che il modello non ha ricevuto.

EvalITA-LLM conferma il quadro e aggiunge un elemento di interesse. La Tabella 10 riporta i sei task a scelta multipla disaggregati per metrica. Emma-6 ottiene 47,2% su textual entailment (8,6 punti sotto Qwen3 e 12,6 sotto Qwen2.5) 20,3% su admission test e 26,1% su FAQ; sulle macro-F1 ottiene 41,5% su sentiment, 39,6% su hate speech e 29,1% su word-in-context.

Due dati meritano attenzione. Il primo: su sentiment Emma-6 ottiene 41,5% di macro-F1, quinta posizione ma davanti a Qwen3 (34,7) e a SmolLM2 (38,7), e a 2,2 punti da Qwen2.5 (43,7). È uno dei pochi task del panel in cui il modello batte Qwen3 con un margine consistente. Il secondo: su word-in-context, che richiede di disambiguare il senso di una parola dal contesto, Emma-6 crolla a 29,1% contro 66,5% di Qwen3: 37 punti di distanza. La spiegazione più plausibile non è la mancanza di conoscenza lessicale (il modello è addestrato su testo Italiano) ma il formato: WiC richiede una decisione binaria su due frasi che condividono una parola, e la rappresentazione di quella decisione dipende dal modo in cui il prompt è costruito, che per Emma-6 non è quello appreso. Il caso è istruttivo perché mostra come lo stesso modello possa essere secondo su un task e ultimo su un altro della stessa suite.

Tabella 10 — EvalITA-LLM: metrica ufficiale per task (%). Le due medie sono calcolate separatamente sulle accuratezze (TE, AT, FAQ) e sulle macro-F1 (HS, SA, WiC), dopo aver mediato le sei varianti di prompt di ciascun task.

Modello	TE (acc)	AT (acc)	FAQ (acc)	HS (F1)	SA (F1)	WiC (F1)	Media acc	Media F1

Modello	TE (acc)	AT (acc)	FAQ (acc)	HS (F1)	SA (F1)	WiC (F1)	Media acc	Media F1
Qwen3 1.7B	55,8	28,9	38,9	60,6	34,7	66,5	41,2	54,0
Qwen2.5 1.5B Instruct	59,8	27,2	47,0	54,4	43,7	52,3	44,7	50,1
Gemma 3 1B IT	56,7	23,3	31,8	62,3	51,3	44,0	37,3	52,5
Llama 3.2 1B Instruct	51,4	23,9	32,6	29,1	44,5	48,1	36,0	40,6
EuroLLM 1.7B Instruct	51,1	26,1	27,6	40,4	43,7	52,2	34,9	45,4
SmolLM2 1.7B Instruct	51,6	20,9	27,4	62,0	38,7	46,8	33,3	49,1
Emma-6	47,2	20,3	26,1	39,6	41,5	29,1	31,2	36,7

ITA-Bench, infine, è il benchmark più completo e più oneroso della batteria: tredici task, sette traduzioni dall'inglese valutate a zero esempi e sei adattamenti da risorse italiane valutati a cinque esempi, per un totale di oltre 30.000 richieste per modello. Emma-6 ottiene 50,5% complessivo, ultima per 0,04 punti dietro SmolLM2 (50,6), con 53,9% sui task tradotti e 46,6% sugli adattamenti. La Tabella 11 riporta il dettaglio.

Tabella 11 — ITA-Bench: accuratezza normalizzata per task (%), valore di Emma-6 e migliore del panel.

Task ITA-Bench	Emma-6	Migliore	Modello
ARC-Easy (tradotto)	42,7	57,0	EuroLLM
BoolQ (tradotto)	58,0	61,5	Qwen2.5
BoolQ con passaggio	45,4	71,6	Qwen2.5
PIQA (tradotto)	63,0	65,1	EuroLLM
SciQ (tradotto)	45,3	59,2	EuroLLM
SciQ con passaggio	65,1	82,3	Qwen2.5
WinoGrande (tradotto)	57,8	61,5	Qwen3
AMI (cloze)	49,9	57,1	Qwen3
GhigliottinAI (cloze)	24,8	36,9	Qwen3
NERMuD (cloze)	54,8	84,8	Qwen2.5
PreLearn (cloze)	48,6	51,2	Qwen3
PreTENS (cloze)	53,1	54,2	Qwen3
WiC (cloze)	48,4	51,2	Qwen2.5

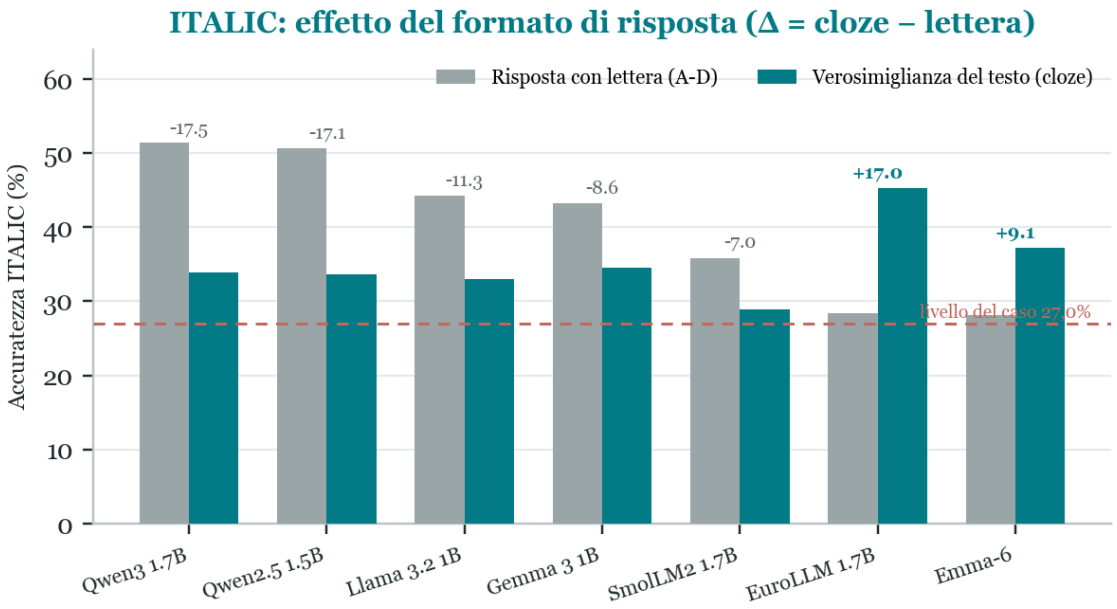
Il dettaglio è più favorevole della media. Su PIQA, che richiede di scegliere la soluzione fisicamente plausibile a un problema quotidiano, Emma-6 è seconda a 2,1 punti dal migliore (63,0 contro 65,1): un risultato che indica comprensione del senso comune applicato, non semplice similarità lessicale. Su PreTENS (53,1 contro 54,2) e PreLearn (48,6 contro 51,2) è a 1,1 e 2,6 punti dal migliore. Il crollo si concentra su due task: NERMuD (54,8 contro 84,8) e GhigliottinAI (24,8 contro 36,9). Il primo richiede di classificare la relazione semantica tra entità in un testo giornalistico (Es: un compito che dipende da un addestramento supervisionato specific). Il secondo è il gioco della ghigliottina, che richiede di individuare la parola legata da un nesso semantico a cinque parole date, ed è notoriamente difficile anche per i modelli grandi (il migliore del panel si ferma al 36,9%).

6.9 Il gap di formato: la diagnosi

Le sezioni precedenti hanno accumulato un'evidenza coerente: Emma-6 è ultima o penultima su MMLU, Belebele, INVALSI, ITALIC e IFEval, ma seconda su bits-per-byte e su TruthfulQA, e seconda su PIQA e PreTENS quando il formato è cloze. L'ipotesi che spiega questa configurazione è che il modello possegga la competenza ma non la politica di risposta richiesta dai protocolli a scelta multipla con etichetta di lettera. L'ipotesi è

verificabile direttamente, perché ITALIC consente entrambe le modalità di valutazione sugli stessi 10.000 item: la modalità a lettera, in cui si confronta la log-verosimiglianza dei token "A", "B", "C", "D", e la modalità cloze, in cui si confronta la log-verosimiglianza del testo dell'opzione.

Figura 8 — ITALIC: accuratezza con risposta a lettera e con verosimiglianza del testo dell'opzione (cloze). Δ positivo indica che il modello migliora quando non deve interpretare un'istruzione.



Il risultato è netto. Emma-6 passa da 28,1% a 37,2% (40,0% con normalizzazione per lunghezza): guadagna 9,1 punti percentuali, il guadagno più ampio del panel. EuroLLM, l'altro modello non allineato, guadagna 17,0 punti (da 28,3 a 45,4). Tutti gli altri perdono: Qwen3 −17,5, Qwen2.5 −17,1, Gemma 3 −8,6, Llama 3.2 −11,3, SmolLM2 −7,0.

La Figura 8 mostra l'inversione: i modelli instruction-tuned sono ottimizzati per emettere un'etichetta quando ricevono un prompt in formato conversazionale, e la modalità cloze li priva di quella politica; Emma-6 ed EuroLLM, che non hanno quella politica, traggono vantaggio da una modalità che chiede solo di continuare un testo plausibile.

Questa è la conclusione tecnicamente più importante del documento, e ha tre conseguenze:

La prima è metodologica: la media aritmetica dei dodici benchmark non è una misura invariante della qualità del modello, perché dipende dal formato di risposta, e un panel che include solo modelli instruction-tuned tenderà sistematicamente a premiare l'allineamento al formato invece della competenza.

La seconda è interpretativa: i punteggi di Emma-6 su MMLU, Belebele, INVALSI e ITALIC non devono essere letti come misure di conoscenza, ma come misure congiunte di conoscenza e di allineamento, e in questi casi la seconda componente domina.

La terza è operativa: esiste un intervento di fine-tuning, di costo contenuto, che produrrebbe un guadagno immediato e misurabile su cinque dei dodici benchmark; è la prima voce della roadmap.

Va detto, per completezza, che il gap di formato non spiega tutto e che la lettura cloze non è quella ufficiale del benchmark. A parità di metrica cloze normalizzata il panel si riordina: EuroLLM 47,8%, Emma-6 40,0%, Gemma 3 37,2%, Qwen3 36,9%, Qwen2.5 36,3%, Llama 3.2 36,2%, SmolLM2 32,0%. In questa lettura Emma-6 è seconda, dietro a EuroLLM e davanti a tutti gli altri. Resta un divario reale di 7,8 punti dal primo, che la sola ipotesi di formato non copre e che va attribuito alla quantità e alla qualità del corpus di addestramento. La conclusione corretta è quindi duplice: la conoscenza culturale italiana di Emma-6 è presente e comparabile a quella dei modelli di pari scala, ma non è pienamente accessibile attraverso i protocolli a risposta chiusa.

6.9.1 Il caso Belebele: anatomia di un punteggio non informativo

Belebele è l'unico benchmark della batteria per cui i dati disponibili consentono di dimostrare, e non soltanto di argomentare, che il punteggio non misura ciò che il benchmark intende misurare. La dashboard della replica indipendente registra, per ogni modello, la distribuzione delle risposte scelte. Incrociandola con la distribuzione delle risposte corrette si ottiene un test diretto dell'ipotesi di collasso posizionale formulata nella sezione 6.3.

Tabella 12 — Distribuzione delle risposte per opzione su Belebele (900 item, %): due modelli a confronto e distribuzione della risposta corretta sullo stesso insieme.

Serie	A	B	C	D
Emma-6	3,9	94,6	1,4	0,1
Qwen3 1.7B	26,6	27,6	25,3	20,6
Risposta corretta	22,9	27,9	27,3	21,9

Emma-6 concentra il 94,6% delle risposte sull'opzione B e dedica lo 0,1% alla D. Qwen3, sugli stessi 900 item, distribuisce le proprie risposte in modo quasi uniforme. Poiché la risposta corretta è la B nel 27,9% dei casi, un modello che rispondesse sempre B otterrebbe 27,9%, cioè 251 risposte corrette su 900: esattamente il totale ottenuto da Emma-6, che ha scelto B in 851 domande. Sulle 49 domande in cui ha scelto un'altra lettera il modello ha guadagnato dieci risposte corrette e ne ha perse dieci, con un saldo nullo. Se si combina la distribuzione effettiva delle risposte con quella delle risposte corrette si ottiene un valore atteso del 27,7%, coerente con il 27,2% misurato nella campagna principale e con il 27,9% della replica indipendente. Il punteggio, in altre parole, è interamente determinato dalla politica di risposta e non contiene informazione sulla comprensione del passaggio. Per stabilire quanto del punteggio dipenda dal formato del prompt servirebbe un confronto che vari il formato a parità di domande, confronto che la campagna non include e che resta una limitazione dichiarata (sezione 8.1).

Il dato resta nella Tabella 6 e nella media a dodici benchmark, perché per gli altri sei modelli il benchmark è informativo e la batteria deve essere comune a tutti: escludere una misura

solo perché penalizza un modello specifico sarebbe una scelta asimmetrica e non difendibile. Ma il 27,2% di Emma-6 non va letto come una misura, pur bassa, di comprensione del testo (è la conseguenza meccanica di una politica di risposta degenere) e la sezione 7.3 ne tiene conto esplicitando un'aggregazione alternativa, calcolata sui benchmark in cui il formato non è in gioco.

6.10 Validazione indipendente

La campagna descritta in questo documento è stata replicata in modo indipendente da Federico Viscioletti, componente della community OpenEmmaLab, con una dashboard di valutazione separata, su macchine diverse (CPU e Apple Metal), con un proprio protocollo di prompting e su snapshot successivi. La replica ha valutato Emma-6 sull'export ONNX FP32, senza token di inizio sequenza, e ha eseguito sia Belebele sia MMLU in regime five-shot (cinque esempi per prompt). Sono tre differenze dichiarate rispetto alla campagna principale, che ha usato il modello in fp16, un token di inizio sequenza determinato empiricamente e cinque esempi per MMLU: vanno tenute presenti nella lettura della Tabella 13, perché concorrono a spiegare l'ordine di grandezza degli scarti osservati. La Tabella 13 confronta i risultati sui benchmark presenti in entrambe le campagne.

Tabella 13 — Confronto tra la campagna di questo documento e la replica indipendente (valori in %, dove presenti due valori: campagna / replica).

Benchmark	Emma-6 (questo doc.)	Emma-6 (indip.)	Qwen3	Qwen2.5	Llama 3.2
Belebele	27,2	27,9	75,1 / 75,4	73,4 / 72,8	44,2 / 44,6
ARC Easy (acc)	42,9	43,0	46,8 / 48,9	43,7 / 45,0	43,1 / 44,5
MMLU anatomia	26,7	26,7	— / 46,7	— / 42,2	— / 37,8

Le differenze sono contenute (tra 0,1 e 1,5 punti) e vanno nella direzione attesa: protocolli di prompting diversi, in particolare nella gestione degli spazi e del formato delle opzioni, producono variazioni di quell'ordine su task a scelta multipla. La convergenza è un elemento di robustezza del dato: il 27% di Emma-6 su Belebele e il 43% su ARC Easy non sono artefatti di una singola implementazione. Nella replica indipendente compare inoltre un dettaglio per singola materia della batteria MMLU, che la campagna principale non ha registrato: sulla materia College physics, 102 item, Qwen3 ottiene 41,2%, Qwen2.5 37,3%, Emma-6 25,5%, EuroLLM 24,5%, e Gemma 3, Llama 3.2 e SmolLM2 si attestano tutte al 21,6%. Emma-6 è quindi terza su sette, non prima: il 25,5% corrisponde a 26 risposte corrette su 102, una in più di EuroLLM e appena mezza risposta sopra le 25,5 attese rispondendo a caso. Letto insieme al collasso posizionale descritto sopra, il dato non è un punto di forza ma la verifica più stringente della diagnosi: su una materia in cui la conoscenza enciclopedica dovrebbe essere discriminante, soltanto Emma-6 (26 risposte) ed EuroLLM (25) restano entro due risposte dal valore atteso per il caso, mentre Gemma 3, Llama 3.2 e SmolLM2, tutte al 21,6% cioè a 22 risposte corrette, se ne allontanano di circa tre risposte e mezza verso il basso, e solo Qwen3 (42) e Qwen2.5 (38) lo superano in modo netto. Va comunque considerato che 102 item valgono circa un punto percentuale ciascuno, quindi la granularità della misura è grossolana e il dato va trattato come osservazione, non come risultato.

La replica indipendente conferma anche il dato qualitativo più importante: la distribuzione delle risposte di Emma-6 su Belebele è fortemente sbilanciata verso l'opzione B (94,6% delle risposte) e non contiene praticamente mai l'opzione D (0,1%). È l'osservazione da cui è partita l'analisi della sezione 6.9, e la sua riproduzione in due campagne distinte rafforza la diagnosi di collasso posizionale rispetto a quella, alternativa, di scarsa comprensione del testo.

6.11 Costo della valutazione

Un'ultima nota di carattere pratico. La batteria completa è costosa: ITA-Bench richiede tra 56 e 88 minuti per modello su una singola GPU da 8 GB, EvalITA tra 12 e 20 minuti, l'intera campagna tra 3 e 6 ore per modello a seconda della dimensione e della presenza di KV-cache. Emma-6 è il modello più lento del panel per token generato, perché l'export ONNX non espone la KV-cache e ogni token richiede un forward completo sulla sequenza: è un limite dell'artefatto di esportazione, non dell'architettura, e la sua rimozione è nella roadmap insieme all'allineamento al formato.

7. Discussione

7.1 Punti di forza

La campagna identifica quattro punti di forza, ciascuno sostenuto da una misura indipendente.

- **Qualità della modellazione linguistica.** Emma-6 è secondo su sette modelli per bits-per-byte su italiano held-out (0,882), davanti a Qwen2.5 (0,939), Llama 3.2 (0,969), Qwen3 (1,037), Gemma 3 (1,145) e SmolLM2 (1,209). Su testi che il modello non ha mai visto, e che sono stati creati dopo la data di addestramento, un modello monolingua da 1,55 miliardi modella l'italiano meglio di modelli più grandi addestrati su corpus cento volte più estesi. È il risultato che più direttamente sostiene la tesi di progetto.
- **Veridicità.** Secondo su TruthfulQA-IT con 46,2% di MC2, a 1,0 punto dal primo e davanti a modelli di dimensioni comparabili addestrati su corpus molto più grandi. La differenza tra MC1 (28,9) e MC2 (46,2) indica che il modello distribuisce la probabilità sulle risposte corrette pur non riuscendo sempre a isolarne una: è il comportamento di un modello prudente, non di un modello che ripete falsi miti.
- **Efficienza computazionale e di esercizio.** Il vocabolario di 64.000 unità calibrato sull'italiano produce 0,217 token per byte contro 0,303 di Qwen2.5 e Qwen3 e 0,354 di SmolLM2: a parità di testo, il 28-39% in meno di token da processare. Unito alla dimensione di 1,55 miliardi e alla disponibilità di un export INT8 da meno di un gigabyte, il risultato è un modello eseguibile su CPU, distribuibile on-premise e sostenibile in esercizio, che è la condizione pratica per l'adozione enterprise e per i contesti con vincoli di sovranità del dato.
- **Riproducibilità e tracciabilità.** La filiera è interamente scriptata, ogni stadio è idempotente, ogni checkpoint porta la propria configurazione, ogni risultato di benchmark è memorizzato in JSON con metrica, revisione del modello, versione

dell'harness e tempi. La valutazione indipendente condotta dalla community ha riprodotto i risultati entro 1,5 punti. Questo è un vantaggio competitivo prima ancora che tecnico: un modello di cui si può mostrare la filiera è un modello che un cliente regolato può adottare.

7.2 Punti di debolezza

I punti di debolezza sono altrettanto netti e non vanno minimizzati.

- **Allineamento al formato di istruzione.** Il modello non dispone di un chat template documentato, e su IFEval-ita ottiene 3,3% contro 14-23% degli altri modelli a parità di prompt nudo. Il deficit si propaga a tutti i benchmark a risposta chiusa in formato few-shot: MMLU 25,0% (esattamente il livello del caso), Belebele 27,2% contro 75,1% del primo, INVALSI 34,3% contro un caso di 32,6%, ITALIC 28,1% contro un caso di 27,0%. Sono numeri che, come si è argomentato, non misurano la conoscenza ma misurano un difetto di interfaccia; restano comunque il principale ostacolo all'uso del modello in applicazioni che richiedono risposte strutturate o scelte tra alternative.
- **Conoscenza enciclopedica in valore assoluto.** Anche al netto dell'effetto formato, il divario rispetto ai modelli multilingua su conoscenza enciclopedica è reale: su ITA-Bench Emma-6 è a 7,4 punti dalla vetta e su ITALIC in modalità cloze normalizzata è a 7,8 punti da EuroLLM. Un corpus di 25 miliardi di token italiani contiene meno sapere fattuale di un corpus multilingua di ordini di grandezza superiore: è un limite fisico dei dati disponibili in lingua italiana, non un difetto di pipeline.
- **Copertura incompleta della batteria.** Dodici dei diciotto task CALAMITA non sono stati eseguiti, i task generativi di EvalITA-LLM sono stati esclusi, la submission alle leaderboard di INVALSI e ITALIC non è stata verificata. La batteria è quindi rappresentativa delle capacità discriminative ma non completa rispetto al brief originale.
- **Assenza di valutazione umana e di valutazione conversazionale.** Tutti i benchmark della batteria sono automatici e a risposta chiusa, con l'eccezione parziale di IFEval. Non è stata condotta alcuna valutazione della qualità conversazionale, della coerenza multi-turno, della sicurezza delle risposte o della qualità della generazione libera. È la lacuna più rilevante per un modello che nasce come assistente, ed è la ragione per cui i risultati di questo documento non devono essere letti come una misura dell'utilità del modello in produzione.

7.3 Perché la media aritmetica non descrive il modello

La media dei dodici benchmark colloca Emma-6 all'ultimo posto con 32,7 contro 47,8 del primo. È un dato vero e va riportato. Non è però una buona sintesi, per tre ragioni tecniche che vale la pena esplicitare perché riguardano qualsiasi confronto tra modelli con e senza allineamento al formato.

La prima ragione è la composizione della batteria. Su dodici benchmark, cinque sono a scelta multipla con etichetta di lettera e richiedono una politica di risposta che Emma-6 non ha: MMLU, Belebele, INVALSI, ITALIC, IFEval. Questi cinque pesano per il 42% della media, e

su di essi Emma-6 prende tra 15 e 27 punti di distacco dalla media degli altri. Un campione sbilanciato verso un formato produce una media che misura quel formato.

La seconda ragione è che i punteggi non sono comparabili in valore assoluto. MMLU su un modello che risponde a caso vale 25%, Belebele su un modello con collasso posizionale vale 27%, IFEval su un modello che non interpreta istruzioni vale 3%. Fare la media di questi numeri con ITA-Bench, dove il caso è il 50%, somma grandezze che non hanno la stessa origine. Una media corretta dovrebbe normalizzare per il livello del caso e per l'errore massimo, ma la normalizzazione non è definita per tutti i benchmark della batteria, e in particolare non lo è per TruthfulQA e per il bits-per-byte.

La terza ragione è che la media premia l'allineamento al formato e non la competenza. Un modello che ha appreso a emettere un'etichetta ottiene vantaggi su cinque benchmark su dodici senza che questo dica nulla sulla qualità del suo italiano. La prova è nell'inversione descritta nella sezione 6.9: passando alla modalità cloze, Emma-6 sale di 9,1 punti su ITALIC e gli altri modelli scendono tra 7 e 17,5 punti.

L'effetto si può quantificare. Ricalcolando la media sui soli sette benchmark il cui protocollo non richiede di interpretare un'istruzione (HellaSwag, ARC Challenge, ARC Easy, TruthfulQA, EvalITA, accuratezza e macro-F1, e ITA-Bench) Emma-6 passa da 32,7 a 39,2, guadagna cioè 6,5 punti e si colloca sesta a 7,3 punti dal primo invece che a 15,1. La seconda aggregazione è riportata in Appendice C, con la sua validazione interna; la prima resta la misura di copertina, perché è quella calcolata su una batteria comune a tutti i modelli.

Il documento riporta quindi la media perché è il dato che il brief richiedeva e perché è il dato confrontabile con la letteratura, ma accompagna ogni affermazione con il dettaglio per benchmark e con la disaggregazione per formato. La conclusione onesta è che Emma-6 è oggi un modello linguistico italiano solido e un assistente non ancora allineato, e che la distanza tra le due cose è colmabile con un intervento circoscritto.

7.4 Implicazioni industriali

Per il mercato italiano il risultato più rilevante non è la posizione in classifica ma la combinazione tra qualità linguistica e costo. Un modello che modella l'italiano meglio di alternative più grandi, che consuma meno token a parità di testo, che pesa un 1,55 miliardi di parametri e che gira in INT8 su CPU è un modello che può essere inserito in un prodotto senza dipendere da un'infrastruttura cloud straniera. Per una pubblica amministrazione, per una banca, per una media company o per un'azienda manifatturiera con documentazione interamente in italiano, questo è il profilo di costo e di conformità che rende l'adozione possibile.

La distinzione rispetto ai modelli generalisti non è quindi di capacità assoluta ma di adeguatezza al dominio: un modello che eccelle su MMLU inglese non è necessariamente il migliore a redigere un bando, a riassumere una delibera o a normalizzare una scheda prodotto in italiano. La campagna non misura queste competenze (è un limite dichiarato) ma misura la condizione che le rende possibili: una buona modellazione della lingua.

8. Limitazioni e lavori futuri

8.1 Limitazioni della valutazione

Le limitazioni della campagna sono di quattro tipi e vanno lette come parte del risultato.

- Copertura: dodici dei diciotto task CALAMITA non sono stati eseguiti; i task generativi di EvalITA-LLM (NER, summarization) sono stati esclusi per costo; la parte leaderboard di ITA-Bench (BBH, MATH, GPQA) non è stata inclusa; la submission alle leaderboard di INVALSI e ITALIC non è stata verificata.
- Statistica: ciascun benchmark è stato eseguito con un solo seed. Gli intervalli di confidenza bootstrap sono riportati nei file di risultato ma non commentati nel testo. Su task con poche centinaia di item, come ARC Challenge (1.169) e TruthfulQA (783), differenze inferiori a 2-3 punti non sono significative.
- Sensibilità al prompt: non è stato condotto uno studio sistematico della varianza rispetto alla formulazione del prompt. La replica indipendente suggerisce variazioni dell'ordine di 1-1,5 punti su task a scelta multipla, ma su IFEval e sui task generativi la varianza può essere molto maggiore. In particolare non è stato eseguito il confronto che servirebbe a separare l'effetto del formato da quello della competenza su Belebele, cioè una variazione del formato del prompt a parità delle stesse 900 domande: i dati di distribuzione delle risposte della sezione 6.9.1 documentano il collasso posizionale, ma non ne quantificano il costo in punti di accuratezza.
- Assenza di valutazione umana: nessuna delle misure riportate riguarda la qualità percepita delle risposte, la coerenza multi-turno, l'utilità o la sicurezza. Il documento non può quindi essere usato per affermare che Emma-6 sia un buon assistente ma solo che è un buon modello linguistico italiano.

8.2 Roadmap tecnica

La roadmap discende direttamente dalle evidenze misurate, in ordine di rapporto tra beneficio atteso e costo dell'intervento.

- Priorità 1 — Chat template e allineamento al formato**.** Definire e pubblicare un chat template ufficiale coerente con il formato appreso in SFT (marcatori "### Istruzione:", "### Input:", "### Risposta:"), e aggiungere al corpus di fine-tuning esempi espliciti di risposta a scelta multipla con etichetta di lettera, di risposta in formato JSON e di esecuzione di vincoli verificabili. Beneficio atteso: guadagno immediato su IFEval, MMLU, Belebele, INVALSI e ITALIC, cioè su cinque dei dodici benchmark, con un intervento di poche ore su un dataset mirato di alcune decine di migliaia di esempi.
- Priorità 2 — KV-cache nell'export ONNX. Il grafo esportato non espone gli stati passati, e la generazione richiede un forward completo per ogni token. È un limite dell'artefatto, non dell'architettura. L'esportazione con input e output di cache ridurrebbe di un ordine di grandezza il costo della generazione e renderebbe competitivo il modello anche su carichi interattivi.
- Priorità 3 — Ampliamento e documentazione del corpus. Il divario di 0,15 bit per byte rispetto a EuroLLM e di 7-8 punti su ITALIC in modalità cloze è un divario di dati. Un

secondo ciclo di addestramento con un corpus più ampio, e con una data card che documenti fonte e licenza per singolo dataset chiuderebbe parte del divario e rafforzerebbe la conformità.

- Priorità 4 — Completamento della batteria e valutazione umana. Eseguire i dodici task CALAMITA mancanti, attivare i task generativi di EvalITA-LLM, verificare la submission alle leaderboard di INVALSI e ITALIC e affiancare alla valutazione automatica una valutazione umana su un campione di prompt reali, con criteri di correttezza, pertinenza, coerenza e sicurezza.
- Priorità 5 — Valutazione della quantizzazione. I benchmark riportati sono stati eseguiti sul modello fp16. Non è stata misurata la perdita di qualità introdotta dall'export INT8, che è il formato destinato alla distribuzione su CPU: una campagna di confronto fp16/INT8 sugli stessi benchmark è necessaria prima di dichiarare equivalenza tra i due artefatti.

9. Conclusioni

Emma-6 è un modello linguistico italiano da 1,55 miliardi di parametri, addestrato da zero su circa 25 miliardi di token italiani di fonti pubbliche, allineato con fine-tuning supervisionato ed esportato in formato eseguibile su CPU. La campagna di valutazione condotta su dodici benchmark italiani con sei modelli open di pari scala produce un profilo coerente. Il modello è secondo su sette per bits-per-byte su italiano mai visto e secondo su TruthfulQA-IT, ma è ultimo nella media generale a causa di un deficit sistematico sui protocolli a scelta multipla in formato few-shot.

L'analisi degli errori identifica la causa di questo deficit con ragionevole certezza. Non la conoscenza, ma il formato. Sugli stessi 10.000 item di ITALIC, Emma-6 guadagna 9,1 punti passando dalla risposta a lettera alla verosimiglianza del testo, ed è l'unico modello del panel, insieme a EuroLLM, a migliorare in quella direzione. Tutti i modelli instruction-tuned perdono tra 7 e 17,5 punti. Su Belebele la distribuzione delle risposte mostra un collasso posizionale (94,6% di risposte B, 0,1% di risposte D) e il punteggio del 27,2% coincide con la frequenza con cui l'opzione B è quella corretta, il che lo rende inutilizzabile come misura di comprensione del testo.

La conclusione tecnica è che Emma-6 è oggi un modello linguistico italiano solido e leggero, con una qualità di modellazione della lingua superiore a quella di modelli più grandi addestrati su corpus prevalentemente inglesi e con un difetto circoscritto di allineamento al formato di istruzione. La conclusione metodologica è che i confronti tra modelli con e senza allineamento al formato richiedono protocolli espliciti e doppie modalità di valutazione. Una classifica costruita su un solo formato misura il formato prima della competenza e questa campagna ne è la dimostrazione empirica.

Il lavoro successivo è definito e di costo contenuto: chat template ufficiale, fine-tuning su formati istruzionali espliciti, esportazione con KV-cache, ampliamento documentato del corpus. I risultati presentati in questo documento costituiscono la misura di riferimento rispetto alla quale valutare quei progressi.

Riferimenti bibliografici

[1] J. Hoffmann et al., “Training Compute-Optimal Large Language Models”, NeurIPS 2022 (arXiv:2203.15556).

[2] B. Zhang, R. Sennrich, “Root Mean Square Layer Normalization”, NeurIPS 2019 (arXiv:1910.07467).

[3] J. Su et al., “RoFormer: Enhanced Transformer with Rotary Position Embedding”, Neurocomputing, 2024 (arXiv:2104.09864).

[4] N. Shazeer, “GLU Variants Improve Transformer”, 2020 (arXiv:2002.05202).

[5] J. Ainslie et al., “GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints”, EMNLP 2023 (arXiv:2305.13245).

[6] T. Kudo, J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”, EMNLP 2018.

[7] R. Rafailov et al., “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”, NeurIPS 2023 (arXiv:2305.18290).

[8] L. Gao et al., “A Framework for Few-Shot Language Model Evaluation”, lm-evaluation-harness, EleutherAI, 2024 (v0.4.13).

[9] R. Zellers et al., “HellaSwag: Can a Machine Really Finish Your Sentence?”, ACL 2019.

[10] P. Clark et al., “Think you have Solved Question Answering? A New Dataset for Elementary Science Questions”, 2018 (arXiv:1803.05457).

[11] D. Hendrycks et al., “Measuring Massive Multitask Language Understanding”, ICLR 2021 (arXiv:2009.03300).

[12] S. Lin, J. Hilton, O. Evans, “TruthfulQA: Measuring How Models Mimic Human Falsehoods”, ACL 2022 (arXiv:2109.07958).

[13] L. Bandarkar et al., “The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Languages”, EMNLP 2023 (arXiv:2308.16884).

[14] Crisp-Unimib, “INVALSI-Eval-Suite: benchmark per la valutazione della comprensione del testo e della riflessione sulla lingua italiana”, 2025.

[15] AILC, “CALAMITA: una suite di valutazione per modelli linguistici italiani”, CLiC-it 2024 (arXiv:2406.17535).

[16] Crisp-Unimib, “ITALIC — Italian Language and Culture benchmark”, 2025, doi:10.5281/zenodo.14725822.

[17] J. Zhou et al., “Instruction-Following Evaluation for Large Language Models”, 2023 (arXiv:2311.07911).

[18] M. Moroni et al., “ITA-Bench: A Benchmark for Italian Language Models”, CLiC-it 2024.

[19] P. Martins et al., “EuroLLM: Multilingual Language Models for Europe”, 2024 (arXiv:2409.16235).

[20] Qwen Team, “Qwen2.5 Technical Report” (2024) e “Qwen3 Technical Report” (2025).

[21] Gemma Team, “Gemma 3 Technical Report”, 2025.

[22] A. Grattafiori et al., “The Llama 3 Herd of Models”, 2024 (arXiv:2407.21783).

[23] L. B. Allal et al., “SmolLM2: When Smol Goes Big — Data-Centric Training of a Small Language Model”, 2025 (arXiv:2502.02737).

[24] EvalITA-LLM, “Valutazione di Large Language Models per l'italiano”, EvalITA 2023-2025.

[25] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act).

Appendice A — Metriche secondarie di Emma-6

Per ogni benchmark sono riportate le metriche secondarie registrate nei file di risultato, utili a distinguere un deficit di conoscenza da un deficit di formato. La colonna sinistra indica la metrica, la destra il valore.

Tabella A1 — Metriche secondarie di Emma-6 sui benchmark della batteria.

Benchmark	Metrica	Valore
EU20-HellaSwag	acc (senza normalizzazione)	34,3
EU20-HellaSwag	acc_norm	41,6
EU20-ARC Challenge	acc	23,4
EU20-ARC Challenge	acc_norm	28,1
ARC Easy	acc	42,9
ARC Easy	acc_norm	40,0
EU20-MMLU	acc 5-shot (caso 25,0)	25,0
EU20-TruthfulQA	acc MC1	28,9
EU20-TruthfulQA	acc MC2	46,2
Belebele	acc 5-shot (caso 25,0)	27,2
INVALSI	acc (caso 32,6)	34,3
INVALSI	riflessione sulla lingua	38,9
INVALSI	comprensione del testo	31,9
ITALIC	acc a lettera (caso 27,0)	28,1
ITALIC	acc cloze	37,2
ITALIC	acc cloze normalizzata	40,0
IFEval-ita	prompt strict acc	3,3
IFEval-ita	instruction strict acc	7,1
Held-out IT	bits per byte	0,882
Held-out IT	byte-perplexity	1,842
Held-out IT	word-perplexity	57,1
Held-out IT	token per byte	0,217
ITA-Bench	media task tradotti (0-shot)	53,9
ITA-Bench	media adattamenti (5-shot)	46,6
EvalITA-LLM	media accuratezze (TE, AT, FAQ)	31,2
EvalITA-LLM	media macro-F1 (HS, SA, WiC)	36,7

Appendice B — Riproducibilità

La campagna è riproducibile a condizione di fissare tre elementi: la revisione dei modelli (Tabella 4), la versione dell'harness (LM Evaluation Harness 0.4.13) e il corpus held-out (impronta MD5 2927b08a, 100 documenti, 261.218 byte). Con questi vincoli, i comandi di esecuzione sono i seguenti.

```
lm_eval --model hf --model_args pretrained=<modello> --tasks hellaswag_it,arc_it,arc_easy_ita,truthfulqa_it_mc1,truthfulqa_it_mc2,belebele_ita_Latn --batch_size auto
```

```
lm_eval --model hf --model_args pretrained=<modello> --tasks m_mmlu_it --num_fewshot 5
```

```
lm_eval --model emma_onnx --model_args model_dir=<emma6>,add_bos=true --tasks m_mmlu_it --num_fewshot 5 --batch_size 8
```

```
lm_eval --model emma_onnx --model_args model_dir=<emma6>,add_bos=true --tasks hellaswag_it --batch_size 8
```

Per Emma-6 l'invocazione avviene tramite il wrapper ONNX registrato nell'harness, che espone i tre metodi richiesti (log-verosimiglianza, log-verosimiglianza su testo, generazione) sopra onnxruntime. I parametri di esecuzione sono: lunghezza massima 4.096 token, lotto 8 sequenze con tetto di 4.096 token per lotto, generazione greedy con massimo 512 token, aggiunta del token di inizio sequenza determinata empiricamente. La replica indipendente descritta nella sezione 6.10 ha invece eseguito Emma-6 sull'export ONNX FP32, senza token di inizio sequenza (add_bos=false), con Belebele e MMLU in regime five-shot: le righe della Tabella B1 si riferiscono alla campagna principale.

Tabella B1 — Parametri di esecuzione della campagna di valutazione.

Elemento	Valore
Harness	LM Evaluation Harness 0.4.13
Seed	1234 (few-shot, bootstrap 1.000 iterazioni)
Lunghezza massima	4.096 token
Generazione	greedy, max 512 token (tutti i modelli)
Precisione Emma-6	fp16 (validato contro fp32, scarto < 1 nat)
Precisione modelli HF	bf16 su GPU, fp32 su CPU
Held-out	100 documenti, 261.218 byte, finestra 3.000 caratteri
Tempo per modello	da 3 a 6 ore (GPU singola, 8 GB)

I dati grezzi della campagna, un file JSON per modello e per benchmark, con metrica primaria, metriche secondarie, tempi, revisione del modello e versione dell'harness — sono conservati insieme al presente documento e costituiscono il riferimento per qualsiasi verifica successiva. La cartella di esportazione contiene inoltre il foglio di calcolo riepilogativo, le figure e il manifest delle revisioni.

Appendice C — Analisi di sensibilità: due aggregazioni

La media a dodici benchmark delle Tabelle 6 e 7 è la misura di copertina, calcolata su tutti i benchmark disponibili per tutti i modelli. La sezione 7.3 argomenta che cinque di essi — MMLU, Belebele, INVALSI, ITALIC e IFEval — richiedono al modello di interpretare un'istruzione e restituire un'etichetta, e che su questa classe Emma-6 non dispone della politica di risposta necessaria: i suoi punteggi si collocano al livello del caso o poco sopra, e la modalità cloze sugli stessi item recupera 9,1 punti (sezione 6.9). La Tabella C1 affianca alla media a dodici una seconda media calcolata sui sette benchmark il cui protocollo non richiede quella politica: HellaSwag, ARC Challenge, ARC Easy, TruthfulQA, EvalITA (accuratezza), EvalITA (macro-F1) e ITA-Bench.

Il criterio è definito dalla classe di formato e non dal segno del risultato: è dichiarato qui una volta per tutte, si applica allo stesso modo a tutti i modelli e non è utilizzato come misura di copertina. La sua validazione è interna al panel: i due modelli che guadagnano di più passando alla seconda aggregazione sono EuroLLM (+7,6 punti) ed Emma-6 (+6,5), cioè esattamente i due classificati come non allineati al formato nella sezione 7.2, mentre Qwen3 e Qwen2.5, i due più allineati, perdono 2,0 e 1,2 punti. Se l'aggregazione premiasse il modello per cui è stata costruita, questo pattern non si osserverebbe.

Tabella C1 — Media a dodici benchmark e media sui sette benchmark a formato compatibile (%, ordinamento per la seconda colonna).

Modello	Media a 12	Media a 7	Differenza
Qwen2.5 1.5B Instruct	47,7	46,5	−1,2
Qwen3 1.7B	47,8	45,9	−2,0
Gemma 3 1B IT	42,1	45,5	+3,4
EuroLLM 1.7B Instruct	37,2	44,8	+7,6
Llama 3.2 1B Instruct	40,1	41,5	+1,4
Emma-6	32,7	39,2	+6,5
SmolLM2 1.7B Instruct	35,7	39,2	+3,5

Nella seconda aggregazione Emma-6 è sesta a 39,2, a cinque centesimi da SmolLM2 (39,2) e a 7,3 punti dal primo. La differenza rispetto a SmolLM2 è ampiamente inferiore all'incertezza della misura e va letta come parità, non come sorpasso. La differenza tra le due medie, 6,5 punti, è la stima del costo del disallineamento di formato e corrisponde all'ordine di grandezza del guadagno ottenuto sulla modalità cloze.