SINTESI DEI CONTENUTI USATI PER L'ADDESTRAMENTO DEL MODELLO EX. ART 53 PAR. 1 LETT D) EU AI ACT

29.05.2025

Il presente documento contiene l'indicazione dei contenuti utilizzati per il training del modello ai sensi dell'art. 53 par. 1 lett d) del Regolamento n. 2024/1689 (EU AI Act), dalla fase di pre-training alla fase di fine-tuning.

Tale documentazione è **resa pubblicamente disponibile** con l'obiettivo di aumentare la trasparenza sui dati utilizzati nelle fasi di pre-addestramento e addestramento del modello di AI per finalità generali, **al fine di agevolare le parti con interessi legittimi** (inclusi i titolari del diritto d'autore e gli interessati) nell'esercitare e far rispettare i loro diritti ai sensi della normativa dell'Unione, rispettando al contempo i segreti industriali.

Tale documentazione deve essere di carattere generale e narrativo, anziché dettagliata sotto il profilo tecnico, deve cioè indicare le principali raccolte o serie di dati che sono state inserite nell'addestramento del modello, quali grandi banche dati o archivi di dati privati o pubblici.

INFORMAZIONI GENERALI			
A.	**Identificazione del fornitore**	Ragione sociale del **fornitore**	Fastweb S.P.A., con sede in Milano, Piazza Olivetti, 1, società a socio unico e soggetta all'attività di coordinamento e di direzione della società Swisscom AG (di seguito, Fastweb").
B.		**Nome** e **Identificatore** univoco della **versione** del modello	FastwebMIIA: Modello Italiano d'Intelligenza Artificiale FastwebMIIA-7B
C.		Base models	N/A
D.		Rappresentante Autorizzato	N/A
E.	**Rilascio**	**Data** di rilascio	29/05/2025
F.	**Dimensione complessiva dei dati di addestramento, modalità e caratteristiche**	**Tipologia dati** di **training**: a) testuali b) immagini c) video d) audio e) altro	Il modello non è multimodale, è stato addestrato utilizzando esclusivamente dati a) testuali per una dimensione complessiva pari a 1.5 * 2 * 10^12 tokens.

G.		Tipologia **dati testuali**: a) testi di narrativa e letteratura b) testi scientifici ed educativi c) notizie, giornalismo e opinioni d) documenti legali e ufficiali e) comunicazione sociale (es. messaggi) f) promozione, pubblicità, recensioni di prodotti e servizi g) altro	a) testi di narrativa e letteratura b) testi scientifici ed educativi c) notizie, giornalismo e opinioni d) documenti legali e ufficiali f) promozione, pubblicità, recensioni di prodotti e servizi g) altri dati tra cui esempi di codici informatici, testi storici e un set di dati sintetici, come da sezioni X e Y.
H.		Tipologia **contenuti grafici**: a) fotografia b) dipinti e belle arti c) infografiche d) illustrazione e design grafico e) immagini tratte da social/personali	Non presenti.
I.		Tipologia **contenuti video**: a) film, spettacolo b) contenuti video animati c) filmati di videogiochi e immersivi (es. 3D) d) documentati e) notizie video e giornalismo f) contenuti degli utenti dei social g) altri contenuti video	Non presenti.
J.		Tipologia **contenuti audio**: a) musica b) narrazione e narrativa (es. audiolibri) c) contenuti audio educativi non di narrativa d) programmi radiofonici e podcast e) comunicazione sociale (telefonate, messaggi vocali) f) altro	Non presenti.
K.		Tipologia **contenuti speciali**: a) codice sorgente b) dati strutturati (es. calendario, mappe, etc.) c) altro, descrivere	a) codice sorgente

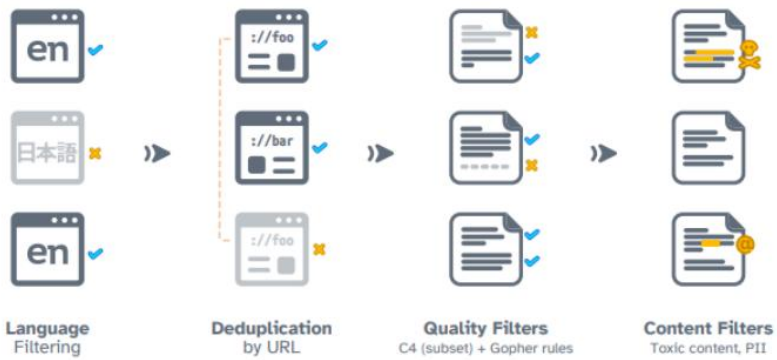
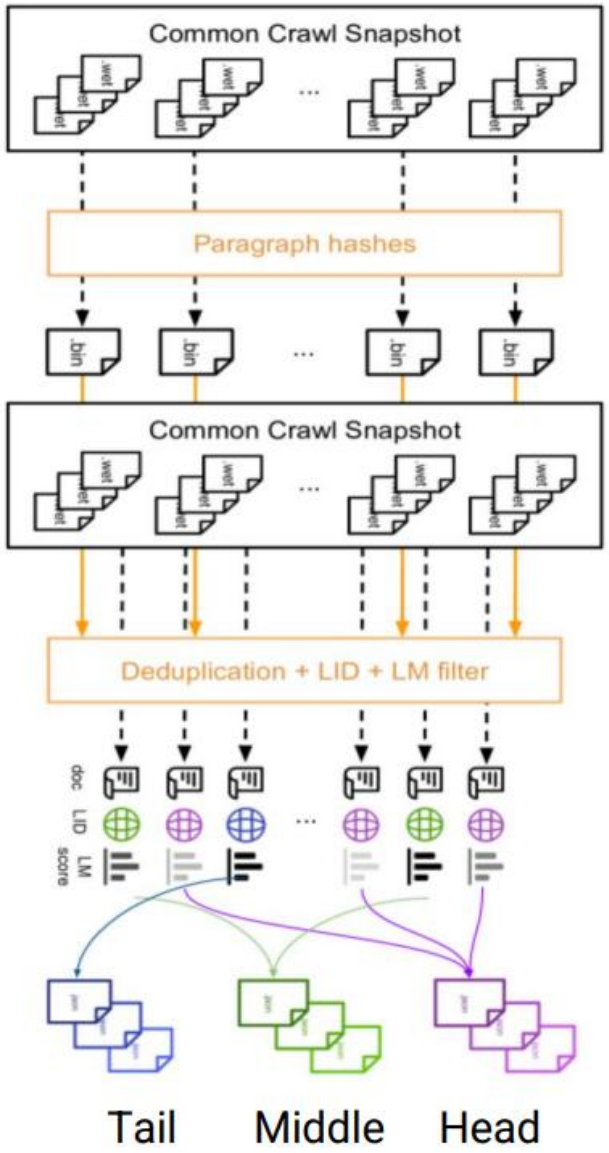
L.		Descrizione delle **caratteristiche linguistiche**, **regionali**, **demografiche** e altre rilevanti dei dati di addestramento complessivi	Il modello è stato addestrato su un corpus eterogeneo di dati esclusivamente testuali, raccolti prevalentemente in lingua italiana e inglese, con una minore presenza di testi in altre lingue europee e non europee. La distribuzione linguistica riflette un focus specifico sull'italiano, al fine di garantire una buona performance del modello su questa lingua. Dal punto di vista del dominio, il dataset copre una vasta gamma di aree tematiche, tra cui: • **Letteratura e scienze umane**: testi narrativi, saggistica, filosofia, critica letteraria; • **Materie scientifiche**: testi relativi a matematica, fisica, biologia, informatica, ingegneria; • **Scrittura di codice informatico**: esempi di codice, principalmente python, documentazione tecnica; • **Nozioni storiche e cultura generale**: contenuti enciclopedici, fonti educative e divulgative; • **Altri ambiti di uso generale**: tra cui diritto, economia, arte, attualità e linguaggio conversazionale; • **Scrittura creativa**: contenuti da quotidiani e articoli divulgativi. Per quanto riguarda la modalità, i dati utilizzati sono esclusivamente testuali. Non sono stati inclusi dati multimodali (es. immagini, audio o video). Tutti questi dati sono stati utilizzati sia in formato 'plain text' durante la fase di pre-training, sia in formato Q&A o multiturn per le fasi di post-training. Per maggiori informazioni vedasi "Documentazione Trasparenza AI" al seguente link.
FONTI DATI			

M.	**Descrizione generale**	Descrizione delle modalità con cui i dati sono stati ottenuti e selezionati	I dati con cui viene addestrato il modello hanno provenienza molteplice. In particolare: 1) Circa l'85% del dataset è tratto da Common Crawl (CC), un archivio realizzato attraverso scraping non indiscriminato sul web, mantenendo la relativa fonte tracciata. Common Crawl, secondo quanto dichiarato dalla stessa organizzazione, rispetta le regole "robots.txt" specificate dai siti web, che governano l'accesso dei web crawler ai loro contenuti e non effettua il crawling di pagine esplicitamente vietate da queste regole; 2) Fastweb, inoltre, ha sottoscritto specifici accordi per utilizzare, a fini di addestramento del modello, i dati di alcuni soggetti terzi con cui ha concluso opportuni accordi di licenza (ad esempio Mondadori, Bignami, Istat, ecc); 3) Una parte minore di dati è stata individuata e acquisita in licenza da fonti aperte (e.g. contenuti di Wikipedia pubblicamente accessibili, acquisiti da libreria su Hugging Face); 4) Una parte inferiore all'1% del corpus di dati è stata generata in modo sintetico tramite software, in modo specifico è stato usato il LLM Phi 3.5 che è rilasciato sotto licenza MIT.
N.	**Dati pubblicamente accessibili**	Dimensione complessiva per modalità e numero di tutti i dataset	Tutte le fonti utilizzate sono state raccolte esclusivamente in modalità testuale. La dimensione supera di poco il 10% del corpus totale.
O.		Elenco dei dataset 'principali/grandi' (oltre il 5% dei dati complessivi in questa categoria) con identificazione univoca, link + periodo di raccolta	Nessun dataset supera il 5% a livello nominale rispetto all'intero corpus. La fonte principale è Wikipedia.
P.	**Dataset privati non accessibili**	Dati concessi in licenza dai titolari dei diritti o dai loro rappresentanti	Tutte le fonti licenziate hanno esclusivamente modalità testuale.
Q.		Dataset acquisiti da altre terze parti	

R.	**al pubblico di terze parti**	Elenco dei dataset privati 'principali/grandi' acquisiti da altre terze parti (oltre il 5% dei dati complessivi in questa categoria), identificatori univoci e link (se disponibili) e descrizione narrativa	Fastweb ha sottoscritto accordi di licenza con partner quali Mondadori, Bignami, Istat.
S.	**Dati raccolti ed estratti da fonti online**	Dimensione complessiva per modalità, periodo di scraping	Circa l'85% del dataset è tratto da Common Crawl (CC), un archivio realizzato attraverso scraping non indiscriminato sul web, mantenendo la relativa fonte tracciata. Common Crawl, secondo quanto dichiarato dalla stessa organizzazione, rispetta le regole "robots.txt" specificate dai siti web, che governano l'accesso dei web crawler ai loro contenuti e non effettua il crawling di pagine esplicitamente vietate da queste regole. Periodo di raccolta: da marzo 2024 a febbraio 2025. Si rinvia a Common Crawl - FAQ, in cui si esplicitano le attività di scraping condotte nel rispetto del protocollo robot.txt, e facendo inoltre il possibile per individuare e rispettare altre riserve espresse sui diritti ai sensi dell'art. 4, par. 3, Direttiva 2019/790.
T.		Identificazione dei crawler, loro scopo e comportamento	Tutte le fonti utilizzate sono state raccolte esclusivamente in modalità testuale. I principali Crawler utilizzati sono: 1) Common Crawl 2) Red Pajama 3) FineWeb Edu
U.		Spiegazione del contenuto che è stato mirato	Sono state predilette fonti contenenti testo di alta qualità in inglese ed italiano. Per la descrizione della tipologia del contenuto si rimanda alla sezione L.
V.		Elenco del 10% dei principali nomi di dominio internet per tipo di modalità dei dati (es. testo, immagine)	1) Common Crawl 2) Red Pajama 3) FineWeb Edu
W.	**Dati forniti dagli utenti (raccolti dal fornitore incl. prompt)**	Dimensione complessiva per modalità	N/A, non vengono utilizzati dati degli utenti
X.		Elenco dei servizi/prodotti del fornitore da cui i dati sono raccolti	N/A, non vengono utilizzati dati degli utenti

Y.	**Dati sintetici auto-generati (dataset)**	Dimensione complessiva per modalità	I dati sintetici costituiscono complessivamente meno del 1% dell'intero corpus.
Z.		Nome del modello di AI	Generati tramite Phi 3.5, rilasciati sotto licenza MIT.
AA.	**Dati acquisiti dal fornitore tramite altri mezzi**	Dimensione complessiva per modalità	N/A
BB.		Mezzi di acquisizione	N/A
ASPETTI RILEVANTI IN MERITO AL TRATTAMENTO DEI DATI			
CC.	**Rispetto del diritto d'autore e dei diritti connessi**	Misure implementate per rispettare le riserve di diritti dall'eccezione di text e data-mining ai sensi dell'Art.4(3) della Direttiva DSM durante la raccolta dei dati, incluse le specifiche dei protocolli di opt-out e delle soluzioni onorate dal fornitore	Fastweb ha utilizzato fonti autorizzate da partnership o dati provenienti da web scraping non indiscriminato che rispettano le regole di 'nofollow' e 'robots.txt'.
DD.		Misure implementate dopo il completamento della raccolta dei dati per identificare e rimuovere i contenuti per i quali i diritti sono stati riservati dai titolari dei diritti	E' stato previsto un canale di comunicazione pubblicamente accessibile per garantire l'esercizio dell'opt-out. Vedasi sezione "Documentazione Trasparenza AI" al seguente link.
EE.	**Rimozione di contenuti indesiderati**	Descrivere i contenuti ritenuti indesiderati dal fornitore come parte dei dati di addestramento	La rimozione dei contenuti indesiderati riguarda un ampio spettro di contenuti che copre sia dati personali (PII), sia valutazioni di qualità dei documenti sia formattazione tipica di dati raccolti dal web. In tutti questi casi si è provveduto all'eliminazione del dato o dell'intero documento di origine.
FF.		Elencare le misure adottate per evitare e/o rimuovere tali contenuti (come blacklist, parole chiave e classificatori basati su modelli)	Fastweb ha creato e tiene aggiornato una blacklist di siti potenzialmente dannosi o contenenti dati personali, per filtrare ulteriormente alla radice alcun urls presenti nel common crawler.
GG.		Le misure applicate dai curatori dei dataset elencati possono essere menzionate, ma non è necessario elencarle in modo esaustivo	È stata sviluppata una pipeline di pulizia dei dati personalizzata e progressiva per incorporare le migliori pratiche presenti in letteratura al fine di generare un corpus italiano progettato per l'addestramento di LLM specifici per la lingua. La pipeline di pulizia prevede fasi di cleaning, normalizzazione, deduplica e filtraggio basato sulla lingua e qualità semantica dei documenti.

			Infine il corpus dati utilizzato è stato più volte bilanciato durante il processo di training per garantire rappresentatività dei diversi domini e copertura dei task desiderati.

HH.		 en 日本語 en ://foo ://bar ://foo Language Filtering Deduplication by URL Quality Filters C4 (subset) + Gopher rules Content Filters Toxic content, PII	 Common Crawl Snapshot Paragraph hashes bin Common Crawl Snapshot Deduplication + LID + LM filter doc LID LM score Tail Middle Head

II.	**Pulizia dati personali**		Occorre chiarire, in primo luogo, il progetto in cui si inserisce la presente valutazione. Per garantire l'efficace funzionamento di un modello LLM è fondamentale che lo stesso sia addestrato con dati testuali, nei quali possono essere anche compresi potenzialmente dati personali. Tali dati, infatti, sono essenziali per garantire un livello adeguato di conoscenza fattuale e contestuale, nonché per migliorare la comprensione e la generazione del linguaggio naturale. Fastweb, occorre precisarlo, non ha interesse a sfruttare direttamente dati personali in chiaro al fine di ottenere un beneficio economico diretto, ma di garantire l'efficacia di un modello linguistico di grandi dimensioni che solo per mezzo dell'addestramento tramite tali dati può risultare performante. Ed invero, senza l'impiego di tali informazioni, il modello risulterebbe limitato nella sua capacità di fornire risposte accurate e contestualizzate, compromettendo la sua utilità e affidabilità. E' per tale ragione che Fastweb ha necessità di trattare alcuni dati personali come, ad esempio, quelli relativi ad eventi storico-scientifici, fatti notori rilevanti sotto il profilo culturale globale, europeo o nazionale. L'addestramento di un modello linguistico di grandi dimensioni (LLM) comporta un equilibrio tra l'esigenza di utilizzare dati testuali, inclusi potenzialmente dati personali, e la necessità di garantire la tutela della privacy. Per ridurre al minimo il rischio per gli interessati, Fastweb ha implementato diverse misure di mitigazione, tra le quali: 1. Selezione accurata delle fonti dati, escludendo dataset contenenti informazioni sensibili o non conformi alle normative sulla privacy. 2. Pipeline di pulizia e anonimizzazione, che rimuove dati personali e informazioni identificabili prima dell'addestramento, assicurando minor impatto e

			presenza dei dati personali. In particolare, per ogni documento, viene applicata un'euristica di rimozione delle Personal Identifiable Information (PII), se presenti in numero inferiore a 5, altrimenti viene rimosso l'intero documento. Si sottolinea, infine, che la principale fonte di dati personali utilizzati per il training è Wikipedia, per la quale valgono le seguenti considerazioni: 1. L'insieme delle versioni utilizzate, in italiano ed inglese, rappresenta solo circa l'1% dell'intero corpus di dati utilizzati; 2. Wikipedia contiene in larga misura informazioni di personaggi storici o famosi (le cui informazioni sono di facile accesso pubblico in ogni caso) o legati a fatti di cronaca esposti mediaticamente (riportate in atti pubblici) 3. la natura principalmente enciclopedica dei contenuti e dei dati personali ivi pubblicati fa propendere per una bassa rischiosità del trattamento degli stessi 4. la pipeline di pulizia per le fonti provenienti da Internet descritta sopra, comprensiva anche di forme di esclusione e masking delle informazioni riferite a persone fisiche identificate o identificabili, viene applicata per tutte le fonti ad eccezione di Wikipedia. Non sono presenti immagini nel dataset usato.