Job Title:

[URGENT] Provide Minimum 200GB Clean Indonesian Pretrained Text Dataset – Multi-Domain NLP (2-Week Delivery)

Job Description:

We are looking for a reliable provider who can deliver a large-scale, high-quality Indonesian pretrained text dataset. The dataset must be clean, ready to use for AI language model training, and legally sourced.

Project Requirements:

Total Volume: Minimum 200GB of clean Indonesian text (pre-collected / pretrained corpus), confirmed by the client upon delivery

Entry Size: Each entry must contain at least 200 words

Domains: accessible domains, including news, encyclopedias, books, technical documents, government reports, forums, or social media (cleaned) etc.

Delivery Timeline: Within 2 weeks of contract start

Data Quality Requirements:

Cleanliness: Text must be fully cleaned and ready to use. No additional cleaning required by the client

Source Transparency: Clear documentation of main data sources and distribution. No un-anonymized personal data

Deduplication: Exact and near-duplicates must be removed (e.g., via Locality Sensitive Hashing). Provide details of deduplication process

Public Dataset Filtering: Avoid content already included in common public datasets (e.g., Common Crawl, OSCAR ,huggingface.co/datasets)

Language Purity: Only Indonesian text allowed. Other languages ≤ 5%

Format and Cleanliness: Must be free of HTML, scripts, ads, navigation elements, OCR errors, duplicates, emojis, symbols, and excessive typos

Data Format Requirements:

Each entry should include:

ID: Unique identifier (UUID)

Language: "id" (Indonesian)

Source URL: Visible and valid

Title: Optional

Text: Cleaned body content

Clean Status: "clean"

Category: Domain label

Folder Organization Example:

lang=id/category=news

lang=id/category=encyclopedia

lang=id/category=book

Who Should Apply:

Fluent in Indonesian

Experienced in large-scale text data delivery (preferably pretrained datasets)

Skilled in text cleaning, deduplication, and multi-domain dataset preparation

Reliable, detail-oriented, and able to meet strict deadlines

Application Instructions:

Submit a short cover letter (English)

State tools, pipelines, or methods used for dataset preparation

Confirm pricing and ability to meet the 200GB / 2-week delivery timeline

We are a serious data company with ongoing multilingual pretrained dataset needs. Top performers may be considered for long-term cooperation.