MainpipeNS Data Pipeline — Summary Report

1. Pipeline Overview

The MainpipeNS pipeline processes raw JSONL text data into a clean, tokenized, packed, and sharded dataset suitable for LLM pretraining.

Stages:

• Inspection & Statistics

• Deduplication

• Cleaning (noise removal, HTML/code filtering, normalization)

• Tokenization with extended GPT■2 tokenizer

• Fixed■length block packing (2048 tokens)

• Train/Val/Test sharding

• Quality analysis (PII, toxicity, perplexity, language distribution)

• Metadata generation

2. Key Quality Metrics (Sample Outputs)

• PII detection: email/phone/credit_card hits counted

• Toxicity (Detoxify): avg toxicity ≈ 0.008 (clean dataset)

• Perplexity (GPT■2 small): median ~34, avg ~45 (high linguistic quality)

• Language distribution: 100% English

• Analysis time: ~200 seconds for 1500 samples

3. Dataset Statistics

• Cleaned dataset reduces noise significantly

• Token length distribution computed pre■packing

• All packed blocks are exactly 2048 tokens

• Sharding ratios: train 98%, val 1%, test 1%

4. Final Output

• tokenized.jsonl — tokenized with BOS/EOS and truncation

• packed_blocks.jsonl — 2048■token blocks padded with <|pad|>

• sharded_dataset/ — split into train/val/test shards

• reports/ — JSON and PDF reports, category distributions

• figures/ — histograms and quality visualizations

• meta.json — tokenizer + dataset metadata

5. Remarks

This pipeline ensures:

• High■quality, English■only data

• Removal of noisy/low■value text

• Safety analysis through PII + toxicity detection

• Reproducible, logged processing stages

• Compatibility with downstream LLM pretraining frameworks