FLM Data and Reproduction Note

From pinned human text and connectome arrays to a verifiable browser model

Kuber Mehta
flm.kuber.studio

12 September 2026 — companion to the FLM working report

Abstract

This note records the data transformations, split boundaries, tokenizer, checkpoint interfaces and verification rules used by FLM. Completed language comparisons use WikiText-2 raw and the BabyLM 2026 10M/100M-word corpora, with vocabularies fitted exclusively on training text. The earlier AMI dialogue study remains separate. Source hashes and machine-readable cards accompany each stage. No raw audio, private speech, pretrained language parameters or teacher-generated training corpus is used. The document distinguishes recorded arithmetic, model inference and independent retraining, and preserves the completed BabyLM evaluation's recovery and generation failures.

1 Dataset choice

The initial question concerned learning from a small amount of human speech. AMI provides human meeting transcripts, but a meeting-only model is narrow and its custom participant-disjoint split is not a standard language-model benchmark. The expanded study therefore adopts WikiText-2 raw [4, 5]: diverse human-written articles with official partitions and an established language-modeling lineage. This changes the domain from spoken conversation to written prose; it must not be described as an improvement in speech understanding.

WikiText-2 is small enough for repeated local training, yet broad enough to expose vocabulary, punctuation, long-document structure and several topic domains. It is not a balanced sample of all human language. The completed larger-data comparison adds BabyLM's standard 10M- and 100M-word corpora [2]. Acquisition, tokenization and overlap audits are complete; their language-training results are a separate experiment and are not included in the WikiText scores.

2 Pinned acquisition and article boundaries

The source is Salesforce's WikiText distribution, configuration `wikitext-2-raw-v1`. The pinned revision is:

```
b08601e04326c79dfdd32d625aee71d232d685c3
```

Three parquet files are downloaded with fixed expected sizes and SHA-256 values. A mismatch fails acquisition. The dataset card stores both source-file and normalized-output hashes, including complete split statistics.

Source row strings are concatenated exactly. Existing newlines are preserved; empty strings add no characters. The pipeline neither lowercases nor invents line breaks. A top-level article heading must use single equals delimiters and be isolated by blank source rows. This second condition matters: a single-equals pattern by itself incorrectly splits some sports-table legends. The fixed parser is tested against that failure and reconstructs exactly the original source concatenation.

State resets occur at article boundaries. Article identities derive from normalized titles for cross-split auditing; separate exact text hashes detect identical documents. The audit found no repeated article

Split	Source rows	Articles	Words	UTF-8 bytes
Train	36,718	600	2,051,910	10,914,845
Validation	3,760	60	213,886	1,144,248
Test	4,358	60	241,211	1,287,656

Table 1: Counts from the exact acquired raw variant. Words are whitespace-delimited counts from this pipeline, not a claim to reproduce another tokenizer's published word counts.

identities or exact duplicated complete articles across partitions. This is not an exhaustive search for overlapping quotations, facts or near-duplicate spans. The raw variant retains preprocessing artifacts and must not be advertised as an untouched Wikipedia source.

Publisher metadata lists CC BY-SA 3.0 and GFDL, while the dataset-card prose links a different Creative Commons version. The FLM card records that discrepancy rather than silently choosing a new license. Downloaded and normalized corpus text remains local and is ignored by Git. The website distributes model parameters, metadata and generated examples, not the source corpus.

3 Lossless training-only vocabulary

The tokenizer uses Hugging Face Tokenizers 0.22.2 [3], byte-level pre-tokenization, no added prefix space and no normalization. Its initial alphabet represents all 256 byte values. BPE learns merges only from the 600 training articles, with a minimum pair frequency of two and 4,094 ordinary vocabulary items. The model reserves two additional integer IDs: BOS=0 and EOS=1.

No special-token strings are registered with the underlying tokenizer. Thus a user literally typing a boundary-looking string is encoded as ordinary text; only the document-building caller inserts boundary IDs. This prevents text content from unexpectedly becoming an instruction to reset or end the model sequence. Unicode, spaces, tabs, newlines and arbitrary byte-alphabet pieces remain representable.

Split	Content tokens	UTF-8 bytes	Bytes/token
Train	3,083,090	10,914,845	3.5402
Validation	322,742	1,144,248	3.5454
Test	367,981	1,287,656	3.4992

Table 2: Frozen-tokenizer encoding statistics. Encoding test text is a deterministic transformation, not test likelihood evaluation or tokenizer fitting.

Every training article round-trips exactly. Each split cache records flat token arrays, document offsets, identities and raw-byte counts. Token lengths reconstruct the likelihood denominator exactly. The browser uses the same byte IDs and ordered merges; independent reference fixtures check exact token IDs across 114 cases including contractions, Hindi, Japanese, emoji, zero-width joiners, unusual spaces, control characters and literal boundary-looking strings.

4 The separate AMI experiment

AMI annotations release 1.6.2 provides manual meeting transcripts under CC BY 4.0 [1]. Participant identities and related meeting families are connected before partitioning, producing 139 training, 16 validation and 16 test documents. Training contains 947,966 whitespace-delimited words and 4,637,005 UTF-8 bytes. Validation and test contain 385,750 and 392,508 bytes respectively.

The AMI path lowercases and linearizes transcript turns with local speaker labels. Overlapping speech becomes an ordered text sequence. These are custom language-model partitions, not the official AMI speech-recognition benchmark. No audio is fed to FLM. The AMI byte vocabulary uses raw byte IDs 0–255 and boundaries 256/257, and its browser checkpoint is identified separately from WikiText. AMI and WikiText adapters cannot be interchanged because their hashes, feature dimensions and vocabularies differ.

5 BabyLM: mixed language at two data scales

All 24 official BabyLM 2026 English files are acquired at immutable repository revisions. They comprise Strict-Small and Strict training plus common development and test partitions. Git blob or LFS identities, source bytes and locally computed SHA-256 hashes are checked. The acquisition totals 706,177,689 bytes. Source components are BNC spoken, CHILDES, Gutenberg, OpenSubtitles, Simple Wikipedia and Switchboard. The training repository metadata advertises MIT, but this does not replace component rights; no raw text is redistributed.

The advertised training budgets correspond exactly to measured whitespace-delimited words: 10,000,000 and 100,000,000. A single 4,096-ID byte-level BPE vocabulary is fitted only to the 10M training partition and reused at both scales. It reconstructs every encoded block's original bytes. No normalization, speaker-marker removal or heading cleanup is applied. The publisher's prior toxicity/bias filtering is distinct from FLM preprocessing and does not establish that the resulting corpus is unbiased.

Partition	Blocks	Text tokens	UTF-8 bytes
10M training	3,351	18,591,514	54,399,840
100M training	33,433	189,709,779	543,302,195
Validation	3,495	19,713,480	56,752,783
Test	3,187	18,177,107	51,722,871

Table 3: Completed BabyLM encoding. Word budgets and model-token counts differ. Boundary IDs are excluded from the text-token column.

The training sources have no blank document separators. Consecutive complete lines are grouped at a 16,384-byte target without crossing source files; oversized individual lines remain intact. Every source byte appears in exactly one block. These computational boundaries are not inferred books or conversations. Each block records source hash, byte offsets, line range, token range and text hash. Tokens use little-endian uint16 memory maps with separate offsets and metadata. Sampling converts only the selected mini-batch to int64. Tests establish identical sampled inputs and likelihoods to an in-memory int64 representation.

The cross-partition audit hashes complete lines after whitespace collapse and casefolding, requiring at least eight words and 40 characters. The 10M training partition matches 289 validation and 1,995 test line occurrences; the 100M partition matches 9,083 and 7,459 respectively. These include potentially common language. Shorter or partial matches and near-duplicates are not covered.

Official partitions remain unchanged for the primary benchmark. A secondary analysis excludes any validation/test block containing a substantial line seen in either training partition. This fixed union excludes 587 validation blocks (9,590,905 bytes) and 659 test blocks (10,759,157 bytes). Whole-block exclusion deliberately avoids splicing new contexts, but changes the domain mixture and can discard much more text than the matched lines themselves. Both architectures and data scales use identical exclusions; neither resulting subset is certified contamination-free.

Before training, eight validation blocks per component are selected by SHA-256 of stable IDs, for a 48-block panel and up to 1,024 scored target positions per block. Seven panel blocks contain audited overlap; selection uses the official panel, with that fact recorded. The compact study assigns 12,000 updates to each of FLM, GRU and transformer at seeds 42 and 43, with the same batch, sequence length, warmup positions and optimizer as WikiText. Its 18,432,000 presented-token budget is about 0.99 times the 10M text-token count and 0.097 times the 100M count, sampled with replacement. This is a fixed-exposure data-diversity comparison, not equal epochs or complete one-pass coverage.

5.1 Completed evaluation, continuations and recovery

All twelve BabyLM fits completed the declared budget. All 288 fixed-panel validation observations reproduce the saved selection: eleven selected checkpoints are at 12,000 updates, and the 100M transformer seed-42 checkpoint is at 11,500. The selection was frozen before held-out likelihood inference. Verification of input identities includes hashing/mapping encoded caches; this is distinct from model scoring. It must not be described as a complete absence of earlier access to test-derived files.

Every selected model is evaluated on all 3,187 official test blocks, retaining per-block likelihood, text tokens,

UTF-8 bytes and source identity. The fixed overlap sensitivity analysis excludes 659 blocks, leaving 2,528 blocks, 13,927,273 targets and 40,963,714 bytes. Results pool negative log likelihood over exact bytes; they do not average component BPBs. State resets between blocks and carries within a block. All three architectures exclude inserted boundary targets at evaluation, while training includes them after warmup.

The first scoring attempt finished one model's block calculations but failed assembling its output because a checkpoint-step keyword was supplied twice. The failed 400-file cache was preserved, the serialization defect corrected and regression-tested, and evaluation restarted from the same frozen weights in a fresh directory because its source identity changed. The corrected supervised sequence completed all twelve scores, all 288 fixed continuations and the separate 64-original-graph cost pilot. The recovery record retains both attempts; training was not repeated for this repair.

The complete output inventory contains twelve original prompts times two sampling seeds times twelve selected models. Every sample reaches its 256-token cap; none terminates at EOS. One invalid UTF-8 sample remains flagged, alongside its token IDs, generated-byte hash and replacement-decoded text. Temperature 0.8, top-k 40 and the explicitly documented display-token filter are shared. Repetition summaries include every output. Neither the prompt categories nor recognizable speaker markers establish conversation, reasoning or factual accuracy.

The published findings link 168 score rows, 84 seed aggregates, all sixteen paired pooled comparisons, per-model reports and all continuations. The table exporter checks full inventory, saved arithmetic, denominators and source identities without restoring weights, reading test-token payloads or repeating the bootstrap. Paper figures are exact copies of visually reviewed, hash-bound PNGs. The source archive supports rebuilding the papers from the included tables and figures; reconstructing inference or fitting requires the separately acquired data and model records. The prior failure and these verification boundaries remain visible rather than being replaced by a successful end-state alone.

6 Connectome and body provenance

The graph snapshot is pinned to revision `776d115ee5aa934578a87fd6d260d138084f59c1` of the public source representation. Acquisition checks 29 required files and reconstructs verified sparse arrays. Counts are 166,700 neurons, 25,582,938 directed neuron-pair edges and 124,177,617 contacts. The complete extraction is distinct from the trained central subset of 1,024 neurons and 76,130 edges.

Selection is based on anatomy before language results. The compact graph retains original neuron IDs, positions, cell-type annotations, edge signs and contact-derived weights. Independent reproduction from the acquired full arrays matches every tracked compact-graph array exactly. Missing soma coordinates remain missing; they are not replaced by invented positions. Anatomical coordinates are locations of somas, not full neuron arbors.

The 39 body meshes retain NeuroMechFly-derived geometry and upstream notices [6]. Automated forward-kinematics checks match the reference joint coordinates. The mesh specimen is female; the graph is male. ChatFLM's body pose summarizes artificial network state. Reproducing the geometry does not reproduce a learned sensorimotor policy, and the language optimizer contains no embodiment objective.

7 WikiText training and selection records

All three main architectures share the same train-only tokenizer, sequence sampler, optimizer schedule and 6,000-update budget. Seeds 42 and 43 produce separate runs. Sampling uses its own RNG, ensuring matched inputs despite different model parameter counts and initialization draws. Checkpoints preserve model and optimizer parameters, sampler RNG state, PyTorch RNG state, exact exposure counts, source revision and tokenizer/graph/cache hashes.

Resume rejects a changed protocol. Tests verify that restoring a checkpoint yields exactly the same next sampled batch and next optimizer update for FLM, GRU and transformer. Validation state resets between articles and streams within each article using native architecture state. Scoring is invariant to article order and chunking within numerical precision. Boundary tokens are excluded; raw token byte lengths determine the denominator.

Final test evaluation requires completion artifacts and unchanged selected-checkpoint hashes for all six

registered main runs. The chosen checkpoints are frozen before the test cache is scored. Results retain per-article negative log likelihood and byte counts. Paired bootstrap comparisons resample matching articles, preserving a common denominator. Intervals describe article-sampling variation conditional on the checkpoints. Two training seeds are insufficient for broad claims about optimizer stability or a general model family.

8 Browser package and local learning

The lexical inference package stores the shared 4,096 by 96 embedding once, input projection, normalized signed recurrent weights in destination-major CSR, pooling assignments, update rates, normalization, output projection and biases. Array shapes, byte offsets, data types and checksums are explicit. The binary is 2,713,236 bytes. Manifest, tokenizer and anatomy hashes are verified before use.

Independent PyTorch fixtures supply full next-token logits, fast and slow states, and projected features after a fixed prefix. JavaScript parity tests check all entries, not a small hand-picked sample. The worker also has integration tests for loading, deterministic generation, scoring, local adaptation, clearing, intervention replay and cancellation during long context priming.

Local learning is an optional readout correction. It uses only user-provided preceding text and next-token targets, with separate before/after probe loss when supplied. The dense correction is exported as little-endian float32 base64 to remain practical for ordinary browser storage; legacy numeric-array exports remain readable. Imports validate checkpoint identity, dimensions, finite values and coefficient bounds. Conversations retain the dataset, checkpoint hash, sampling settings and intervention settings used for each generated passage.

9 Reproduction entry points

The repository README provides the complete current command sequence. The core entry points are:

Command	Purpose
`python -m flm.wikitext`	Pinned text acquisition and article audit
`python -m flm.tokenizer`	Train-only vocabulary and frozen split encoding
`python -m flm.language_suite`	Registered main training runs
`python -m flm.language_report`	Validation artifact consolidation
`python -m flm.language_test`	Gated final test evaluation
`python -m flm.export_lexical`	Browser export; checkpoint argument required
`npm test`	Browser-core, worker and geometry tests
`npm run build`	Static production build and attribution pages

Raw corpora, complete graph downloads and training intermediates stay in ignored local directories. Compact graph assets, tokenizer metadata, browser packages, reports and meaningful code changes are versioned in sequential commits. GitHub Pages hosts the generated public interface at https://flm.kuber.studio. Repository publication is authorized for 12 September 2026; the site and standalone bundles are available independently of a checkout. No API secret or remote language service is needed for browser inference.

WikiText and BabyLM comparisons are complete. The active 128-fit native-Mac selection study retains separate provenance; its validation and held-out comparisons remain pending and may follow repository publication. Broader scaling, alternative learning and learned sensorimotor transfer are deferred. Each stage retains separate provenance and evaluation. Reproducibility is evidence that the specified computation was performed; it is not evidence that the computation reproduces an animal's cognition.

References

[1] AMI Consortium. AMI Meeting Corpus: annotations and documentation.

https://groups.inf.ed.ac.uk/ami/corpus/, 2026. Annotations release 1.6.2, accessed 10 September 2026.

[2] BabyLM Challenge organizers. Babylm 2026 guidelines. https://babylm.github.io/guidelines.html, 2026.

[3] Hugging Face. Tokenizers: byte-level pre-tokenization documentation. https://huggingface.co/docs/tokenizers/api/pre-tokenizers, 2026. Implementation version 0.22.2.

[4] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. https://arxiv.org/abs/1609.07843, 2016.

[5] Salesforce Research. WikiText dataset card and pinned raw partitions. https://huggingface.co/datasets/Salesforce/wikitext, 2026. Revision b08601e04326c79dfdd32d625aee71d232d685c3, accessed 10 September 2026.

[6] S. Wang-Chen et al. NeuroMechFly v2: simulating embodied sensorimotor control in adult Drosophila. *Nature Methods*, 21:2353–2362, 2024. https://doi.org/10.1038/s41592-024-02497-y.