

Public Summary of Training Content — MOTHER CORE V2 & V3

A sufficiently detailed summary of the content used to train the general-purpose AI models MOTHER CORE V2 and V3, prepared according to the template provided by the European AI Office.

PROVIDER

Media Stream AI Limited (MSAI), United Kingdom

MODELS

MOTHER CORE V2 · MOTHER CORE V3
(open-weight, Hugging Face)

LEGAL BASIS

Reg. (EU) 2024/1689 Art. 53(1)(d) & AI Office template

BASIS OF FIGURES

Measured from the training corpus (1,662,215 records); shares by record count

§1 General information

1.1 Provider & contact

Media Stream AI Limited (MSAI), United Kingdom. Point of contact for this summary: ai-governance@mediastreamai.com.

1.2 Model identity

MOTHER CORE V2 and V3 are the same custom architecture at successive checkpoints of one continued training run.

ATTRIBUTE	MOTHER CORE V2	MOTHER CORE V3
Modalities (input → output)	Text → text	Text → text
Release form	Open weights (Hugging Face)	Open weights (Hugging Face)
Architecture	Custom MotherCoreModel (decoder-only)	Custom MotherCoreModel (decoder-only)
Parameters	~6.9B	~6.9B

ATTRIBUTE	MOTHER CORE V2	MOTHER CORE V3
Layers / hidden / attention	48 · 3072 · 24/6 GQA · SwiGLU	48 · 3072 · 24/6 GQA · SwiGLU
Position / norm / context	RoPE $\theta=10000$ · RMSNorm · 4096 · bf16	RoPE $\theta=10000$ · RMSNorm · 4096 · bf16
Tokenizer / vocab	SentencePiece · 50258	SentencePiece · 50258
Training-data cut-off	[cut-off date]	[cut-off date]

1.3 Training corpus

V2 and V3 were trained on MSAI’s Phase-One curriculum — a balanced-sampled agentic corpus of 2,400,092 training records drawn from a source corpus of **1,662,215 unique records**. The data-source figures in §2 are measured over the unique-record corpus.

§2 Data sources used for training

Data sources by category of origin, measured over the 1,662,215-record corpus. Shares are by record count.

CATEGORY	RECORDS	SHARE
2.1 Publicly available datasets	852,670	51.30%
2.2 Licensed / third-party private datasets	0	0.00%
2.3 Data crawled or scraped from the web (first-party)	0	0.00%
2.4 User data	0	0.00%
2.5 Synthetic data (MSAI-generated)	461,705	27.78%
2.6 Other sources (internally created / contributed)	342,840	20.63%
2.7 Other web-derived sample	5,000	0.30%

2.1 Publicly available datasets — 51.30%

Openly available datasets obtained from Hugging Face and equivalent repositories, grouped by domain:

- **Instruction & chat:** allenai/tulu-3-sft-mixture, allenai/WildChat-1M, Magpie-Align/Magpie-Pro-300K-

- Filtered, OpenAssistant/oasst2, databricks/databricks-dolly-15k, nvidia/HelpSteer2, euclaise/writingprompts.
- **Question answering & knowledge:** microsoft/ms_marco, mandarjoshi/trivia_qa, hotpotqa/hotpot_qa, google-research-datasets/nq_open, rajpurkar/squad_v2, stanfordnlp/web_questions, microsoft/wiki_qa, allenai/ai2_arc, allenai/sciq, derek-thomas/ScienceQA, cais/mmlu & TIGER-Lab/MMLU-Pro, allenai/winogrande, OpenOrca/OpenOrca, Wikipedia (English).
 - **Mathematics & reasoning:** meta-math/MetaMathQA, AI-MO/NuminaMath, openai/gsm8k, open-r1/OpenR1-Math-220k, nvidia/OpenMathReasoning, ServiceNow-AI/R1-Distill-SFT, open-thoughts/OpenThoughts-114k, glaiveai/reasoning-v1-20m, bethgelab/CuratedThoughts, bespokelabs/Bespoke-Stratos-17k, simplescaling/slK, openbmb/UltraData-SFT-2605. Certain of these reasoning datasets consist of model-distilled reasoning traces; they are obtained and counted as publicly available datasets.
 - **Code:** glaiveai/glaive-code-assistant, ise-uiuc/Magicoder-OSS-Instruct-75K & Magicoder-Evol-Instruct-110K, theblackcat102/evol-codealpaca-v1, m-a-p/CodeFeedback-Filtered-Instruction, nvidia/OpenCodeReasoning, KodCode/KodCode-V1, bigcode/self-oss-instruct-sc2-exec-filter-50k, sahil2801/CodeAlpaca-20k, jtatman/python-code-dataset-500k, christopher/rosetta-code, google/code_x_glue_cc_defect_detection, jcordon5/cybersecurity-rules.
 - **Summarisation:** abisee/cnn_dailymail.
 - **Safety / alignment:** PKU-Alignment/PKU-SafeRLHF, PKU-Alignment/BeaverTails, Anthropic/hh-rlhf.
 - **Multilingual / translation:** CohereForAI/aya; OPUS parallel corpora for Welsh, Irish and Scottish Gaelic.

2.2 Licensed / third-party private datasets — none

No licensed or privately-licensed third-party dataset was used to train these models.

2.3 Data crawled or scraped from the web (first-party) — none

No first-party web crawling was carried out for these models. Web-origin content was obtained only as part of the published third-party datasets in §2.1.

2.4 User data — none

No platform or user data was used to train these models.

2.5 Synthetic data (MSAI-generated) — 27.78%

Data generated by MSAI rather than obtained from a third party — the second-largest category, reflecting the agentic focus of the curriculum:

- **Agentic tool-use & retrieval scenarios:** generated MOTHER agent traces — single and multi-step tool-calling, planning and decomposition, and retrieval-augmented (RAG) single-call and synthesis scenarios.
- **Agentic-scenario generators:** generated tool-routing, DAG-orchestration, error-recovery, grounding/citation and delegation scenarios.
- **Procedural & curriculum generation:** generated arithmetic and calculator curricula and generated reasoning material.

2.6 Other sources (internally created / contributed) — 20.63%

Datasets authored or curated by MSAI (hand-authored, curated, or assembled from open reference material):

- **MOTHER Defence & safety-critical:** observe-and-advise operator scenarios and threat reasoning (non-kinetic, human-in-the-loop).
- **Security / cyber (defensive):** defensive cyber, red-team, vulnerability and detection-rule material (the latter assembled from open-source detection rules), used to teach defensive posture and refusal.
- **Partner agent training:** MSAI-authored agent and vertical training for the AUTM platform.
- **UK, identity & language curation:** hand-curated UK core facts and model-identity data, and Welsh, Irish and Scottish-Gaelic curation.
- **Assembled mathematics:** an MSAI-assembled mathematics set.

- **MSAI archive:** earlier MSAI-produced question-answer, knowledge and explanation sets.

2.7 Other web-derived sample — 0.30%

A small news-derived question-answer sample of approximately 5,000 records drawn from BBC news content. This was ingested as a prepared sample, not obtained by first-party crawling.

§3 Relevant data-processing aspects

3.1 Respect for rights reservations (TDM opt-outs)

MSAI applies the measures set out in its EU Copyright Compliance Policy (Art. 53(1)(c), companion document): honouring robots.txt, the TDM Reservation Protocol (TDMRep), AI-usage declarations, and per-work licence reservations; exclusion of reserved content at acquisition and dataset-assembly time; operation of an identifiable crawler; and a rights-holder removal channel. The publicly available datasets in §2.1 are used under the terms on which they are published.

3.2 Measures against illegal content & data quality

The training corpus is assembled predominantly from established, curated public datasets and from MSAI-generated and internally curated data; MSAI does not perform indiscriminate first-party web scraping. Before training, MSAI applies:

- **Content cleaning** — removal of extraction artefacts and degenerate or malformed text.
- **Deduplication** — near-duplicate and exact-duplicate removal across the corpus.
- **Quality & format auditing** — automated per-category format and trainability audits, and pre-training corpus gates that reject malformed records.
- **Safety & alignment data** — inclusion of safety/alignment datasets (§2.1) and MSAI refusal/observe-and-advise material to train safe behaviour and refusal.

MOTHER models are developed under a non-weapon, human-in-the-loop posture (observation and decision-support only).

§4 Notes on measurement

The figures in §2 are produced by aggregating each training record's recorded source label and category across the corpus and classifying each source into the categories above. 99.7% of records are attributed to a named category; the residual is the small news-derived sample in §2.7. A more granular reference measurement over MSAI's later, expanded corpus (which did not train the published V2/V3) is retained internally.

MSAI · MOTHER — Public Summary of Training Content · Art. 53(1)(d), Regulation (EU) 2024/1689, prepared to the European AI Office template. Media Stream AI Limited, United Kingdom. Figures measured from the training corpus (1,662,215 records; shares by record count).