# Public Summary of Training Content for General-Purpose AI Models

Mellum 2 model family public summary for the listed Mellum 2 model weights.

| | |
|---|---|
| **Version of the Summary** | v1.0 |
| **Last update** | 29/05/2026 |
| **Scope** | Mellum 2 model family. |
| **Latest date of data acquisition/collection for model training** | 5/2026 |

## 1. General information

### 1.1 Provider identification

| | |
|---|---|
| **Provider name and contact details** | JetBrains s.r.o.<br>mellum@jetbrains.com |
| **Authorised representative name and contact details** | Not applicable. |

### 1.2 Model identification

| | |
|---|---|
| **Versioned model name(s)** | • JetBrains/Mellum2-12B-A2.5B-Thinking<br>• JetBrains/Mellum2-12B-A2.5B-Instruct<br>• JetBrains/Mellum2-12B-A2.5B-Thinking-SFT<br>• JetBrains/Mellum2-12B-A2.5B-Instruct-SFT<br>• JetBrains/Mellum2-12B-A2.5B-Base<br>• JetBrains/Mellum2-12B-A2.5B-Base-Pretrain |
| **Model dependencies** | • Base-Pretrain: pre-trained base.<br>• Base: Base-Pretrain with 128K context extension.<br>• Instruct-SFT: Base + supervised fine-tuning.<br>• Thinking-SFT: Base + supervised fine-tuning with reasoning traces.<br>• Instruct: Instruct-SFT + RLVR.<br>• Thinking: Thinking-SFT + RLVR. |
| **Date of placement of the model on the Union market** | 29/05/2026 |

Scope: Covers the listed Base, SFT, Instruct, and Thinking variants.

### 1.3 Modalities, overall training data size and other characteristics

Modality

Training data size

Types of content

[x] Text

Selected range: More than 10 trillion tokens.

Pre-training: ~10.65T tokens.

Additional training: ~117B long-context tokens, ~47B Instruct SFT tokens, ~167B Thinking SFT tokens, and RLVR for final Instruct/Thinking variants.

Web and general knowledge text; source code; web pages containing code; mathematical text; educational web content and PDFs; multilingual reasoning and QA; curated SFT/STEM/knowledge sources; Wikipedia rewrites; synthetic encyclopedic articles; synthetic and derived code annotations; chat and instruction-following; tool use/function calling; agentic coding; safety and identity examples.

[ ] Image

Not applicable.

Not applicable.

[ ] Audio

Not applicable.

Not applicable.

[ ] Video

Not applicable.

Not applicable.

[ ] Other

Not applicable.

Programming language source code is treated as text.

**Description of the linguistic characteristics of the overall training data**

Multilingual text and QA data. Source code covers Python, Java, PHP, TypeScript, C#, JavaScript, JSX, Rust, Kotlin, Go, C++, and CSS. Detailed language shares are not specified.

**Other relevant characteristics of the overall training data**

Software-engineering focus. Training shifts toward code and math:

- Phase 1: 70% web, 23% code, 6% math.
- Phase 2: 44% web, 42% code, 14% math.
- Phase 3: 23% web, 59% code, 18% math.

Code data includes permissively licensed public repositories, deduplicated at file level. Long-context training includes repository-level FIM examples.

**Additional comments**

Vocabulary: 98,304 tokens. Objectives: next-token prediction and fill-in-the-middle.

## 2. List of data sources

### 2.1 Publicly available datasets

| | |
|---|---|
| **Have publicly available datasets been used?** | [x] Yes [ ] No |
| **Modality of content covered** | [x] Text [ ] Image [ ] Video [ ] Audio [ ] Other |

| | |
|---|---|
| **List of large publicly available datasets** | Large/source categories include:<br>• Common Crawl-derived web corpora and code pages.<br>• Public repositories with permissively licensed source code.<br>• Educational web content and PDFs.<br>• Multilingual reasoning and QA datasets.<br>• Curated knowledge, SFT, STEM instruction, Wikipedia rewrites, and synthetic encyclopedic articles.<br>• Math web content, permissively licensed math textbooks, math SFT, and math instruction data.<br>• Public RLVR/SFT sources including OLMo-3 math RL, Nemotron/NeMo Gym-style math, tool-use and instruction tasks, reasoning-gym, xLAM-style function-calling, and workplace-assistant tasks.<br>The approximate start date of the data collection is not known. The data collection ended in May 2026. |
| **General description of other publicly available datasets not listed above** | Other public text sources include code, instruction data, QA, STEM material, math tasks, structured-output tasks, calendar-scheduling tasks, tool-use tasks, and procedurally generated reasoning tasks. |
| **Additional comments** | Dataset-level sizes, dates, and links are not specified. |

### 2.2 Private non-publicly available datasets obtained from third parties

#### 2.2.1 Datasets commercially licensed by rightsholders or their representatives

| | |
|---|---|
| **Have transactional commercial licensing agreements been concluded?** | [ ] Yes [x] No |
| **Modality of content covered** | Not applicable. |

#### 2.2.2 Private datasets obtained from other third parties

| | |
|---|---|
| **Have private datasets been obtained from other third parties?** | [ ] Yes [x] No |
| **Modality of content covered** | Not applicable. |
| **If publicly known, list private datasets from other third parties** | Not applicable. |
| **General description of non-publicly known private datasets** | Not applicable. |
| **Additional comments** | Provider-created sources are listed in Section 2.6. |

### 2.3 Data crawled and scraped from online sources

| | |
|---|---|
| **Were crawlers used by the provider or on behalf of the provider?** | [ ] Yes [x] No |
| **Crawler name(s)/identifier(s)** | Not applicable. |
| **Purposes of crawler(s)** | Not applicable. |
| **General description of crawler behaviour** | Not applicable. |
| **Period of data collection** | Not applicable. |
| **Comprehensive description of type of content and online sources crawled** | Not applicable. |

| | |
|---|---|
| **Type of modality covered** | Not applicable. |
| **Summary of most relevant domain names crawled** | Not applicable. |
| **Additional comments** | Common Crawl-derived data is listed in Section 2.1. |
| **Have sources other than Sections 2.1 to 2.5 been used?** | [x] Yes |
| **Narrative description of these sources and data** | Provider-created and task-specific sources include:<br>• Multi-language coding tasks: Python, Java, PHP, TypeScript, C#, JavaScript, JSX, Rust, Kotlin, Go, C++, and CSS.<br>• Greenfield implementation, debugging, test generation, behavior modification, filesystem/API integration, and security hardening tasks.<br>• Agentic coding trajectories and SWE-style repository-level edit tasks.<br>• Tool-use and function-calling conversations. |

#### 2.4 User data

| | |
|---|---|
| **Was data from user interactions with the AI model used to train the model?** | [ ] Yes [x] No |
| **Was data from user interactions with the provider's other services or products used?** | [ ] Yes [x] No |
| **Description of provider services or products used to collect user data** | Not applicable. |
| **Type of modality covered** | Not applicable. |
| **Additional comments** | Not applicable. |

#### 2.5 Synthetic data

| | |
|---|---|
| **Was synthetic AI-generated data created by or on behalf of the provider used?** | [x] Yes [ ] No |
| **Modality of synthetic data** | [x] Text [ ] Image [ ] Video [ ] Audio [ ] Other |
| **General-purpose AI model(s) used to generate synthetic data** | Not specified. |
| **Other AI models, including provider-owned models not available on the market** | Not specified. |
| **Additional comments** | Synthetic/derived text includes code QA, code rewrit |

#### 2.6 Other sources of data

| | |
|---|---|
| | • Safety refusal/safe-response data and identity examples.<br>• Long-context repository-level FIM examples. |
| **Additional comments** | Some sources may overlap with public or synthetic categories. |

## 3. Data processing aspects

#### 3.1 Respect of reservation of rights from text and data mining exception or limitation

| | |
|---|---|
| **Signatory to the Code of Practice for GPAI models?** | No |

| | |
|---|---|
| **Measures implemented before model training to respect reservations of rights** | Raw code uses mostly permissively licensed public rep |
| **Additional comments** | N/A |

### 3.2 Removal of illegal content

| | |
|---|---|
| **General description of measures taken** | Training data is obtained from sources with low risk of illegal content selected individually b |

### 3.3 Other information (optional)

| | |
|---|---|
| **Other relevant data processing information** | • Three-phase web/code/math curriculum.<br>• File-level code deduplication and intra-phase web deduplication.<br>• Dataset repetition capped at 4x.<br>• Fill-in-the-middle examples with PSM/SPM ordering.<br>• Best-fit packing to 8,192-token pre-training and 131,072-token SFT sequences.<br>• Long-context extension with layer-selective YaRN.<br>• RLVR with code, math, judge, and task-specific verifiers. |

Appendix: variant-specific training stages

| Model identifier | Training stages and approximate extra data |
|---|---|
| JetBrains/Mellum2-12B-A2.5B-Base-Pretrain | Pre-training on ~10.65T tokens; native context length 8,192. |
| JetBrains/Mellum2-12B-A2.5B-Base | Base-Pretrain plus long-context extension to 131,072 tokens; ~117B long-context trai |
| JetBrains/Mellum2-12B-A2.5B-Instruct-SFT | Base plus supervised fine-tuning as a no-thinking assistant; ~47B SFT tokens. |
| JetBrains/Mellum2-12B-A2.5B-Thinking-SFT | Base plus supervised fine-tuning as a reasoning-trace assistant; ~167B SFT tokens. |
| JetBrains/Mellum2-12B-A2.5B-Instruct | Instruct-SFT plus RLVR on the Instruct mix: ~260,500 training prompts and ~3,600 va |
| JetBrains/Mellum2-12B-A2.5B-Thinking | Thinking-SFT plus RLVR on the Thinking mix: ~259,000 training prompts and ~3,600 |