MIMIR: A SUB-2B HIERARCHICAL REASONING MODEL TRAINED ON PERMISSIBLE DATA PUNCHES ABOVE ITS WEIGHT CLASS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training capable language models typically relies on web-scale pretraining followed by multi-stage post-training, creating substantial computational and data-governance barriers for settings in which training data must satisfy explicit permissibility constraints. We introduce Mimir, a 1.78B-parameter language model based on the Hierarchical Reasoning Model Text (HRM-Text) architecture, a recurrent-depth transformer whose layers are split into a low-level (L) and a high-level (H) stack that are applied repeatedly in an interleaved schedule. Mimir is trained from random initialization directly on a curated mixture of instruction, reasoning, tool-use, multilingual, and synthetic data: 161 datasets and 70.5B tokens per epoch for Mimir v1, extended to 103B tokens per epoch for its continued-training successor Mimir v1.5, with synthetic replacements for sources that do not satisfy permissibility criteria. Across 20 benchmarks in English, Danish (a low-resource language with $\sim$6M speakers), and Math & Code, Mimir v1.5 attains the highest average score of all sub-2B models in every domain, outperforming both HRM-Text 1B and the looped transformers Ouro 1.4B and Ouro 2.6B, and Mimir matches or exceeds the performance of models of up to 9B parameters, such as Qwen 3.5 9B and Gemma 4 E4B, on the English and Danish averages. Two independent memorisation audits indicate low memorisation risk despite multi-epoch training. Inference-time ablations show that the number of L-stack applications can be changed with graceful degradation, whereas changing the number of H-stack applications breaks the model. Our results provide evidence that hierarchical recurrent language models can learn effectively from post-training-oriented, permissible data mixtures without conventional web-scale pretraining.

1 INTRODUCTION

Current developments in LLM training are largely driven by a monolithic recipe consisting of massive, multi-stage pipelines trained on exorbitant volumes of data (Wang et al., 2026). This recipe requires vast compute for pre-training and creates a prohibitive entry barrier for the broader community of researchers and practitioners. This barrier is especially high for efforts, including several in Europe, that commit to regulation, copyright law, and the use of exclusively permissible and ideally openly licensed data: the pool of such high-quality data is limited for most languages, and even where data exists, pre-, mid-, and post-training demand massive compute. Training capable, instruction-ready language models from scratch under these constraints therefore remains a key challenge.

Tackling this challenge, we present **Mimir**, which combines the HRM-Text architecture (Wang et al., 2026) with scalable instruction data pipelines for training. The recurrent depth architecture, along with a PrefixLM attention mask enable focusing on post-training data from the beginning of training. As a result, Mimir is a 1.78B hierarchical reasoning model, trained from scratch, that leads the Pareto frontier on both English and Danish, outperforms similar recurrent depth models, and competes with larger models of up to 9B parameters (Figure 1).

In more detail, Mimir v1 is trained on a curated mixture of 161 English and Danish datasets comprising approximately 70.5B tokens per epoch, a five-fold increase over the training data of HRM-Text

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1B. Furthermore, as certain datasets used in (Wang et al., 2026) do not align with our permissibility standards, we demonstrate the efficacy of synthetically generating "transplant" datasets, replacing non-permissible data with synthetically generated permissible variants. With module-specific learning rate scaling as well as expansion to 103B tokens per epoch, iterative auditing, and refinement of the datasets, we continually train the model to attain Mimir v1.5. Mimir v1.5 is best-in-class amongst recurrent depth approaches and competes with larger frontier models of sizes up to 9B.

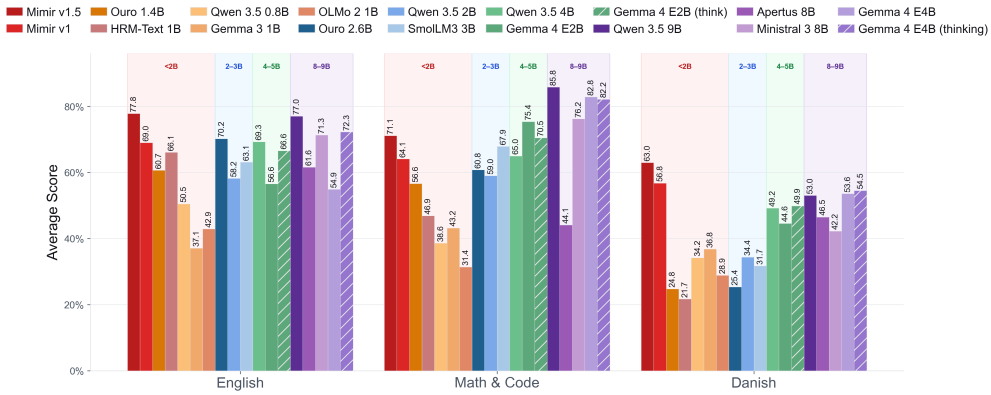


Figure 1: Aggregate results across three categories English – Math & Code, and Danish – comparing Mimir v1 and v1.5 against frontier models from 0.8B to 9B, such as HRM-Text 1B, Gemma 4 E4B, and Qwen 3.5 9B, displaying highly competitive performance across 20 benchmarks against both frontier models and leading European private and public efforts.

The core of our contribution is a set of synthetic instruction-data pipelines with automated auditing, from which we assemble the Mimir Training Corpus (70.5B tokens per epoch for v1 and 103B for its continued-training successor v1.5). We combine HRM-Text with the Gemma 4 tokenizer and chat template, and gradually extend backpropagation through the recurrence over the course of training.

Our results show that our synthetic data pipeline leads to comparable or superior performance without compromising data rights, further underscoring the relevance of HRMs for low-resource language modeling, empowering communities of practitioners and researchers with small, capable and fully permissible models, with both low training and inference budgets.

We evaluate Mimir on 20 standard benchmarks in English, Math & Code, and Danish. Two independent memorisation audits find that the fraction of memorised responses is negligible, indicating low copyright risk under the tested scenarios. Finally, inference-time ablations of the recurrence show that the number of low-level (L) module applications is elastic, whereas the number of high-level (H) applications is not.

In sum, our contributions are as follows:

- We introduce **Mimir** (§3), an HRM-based language model trained from random initialization directly on a permissible mixture of instruction, reasoning, multilingual, tool-use, and synthetic data (161 datasets, 70.5B sampled tokens per epoch for Mimir v1), together with Mimir v1.5, which has been continually trained on 103B tokens per epoch.
- We introduce a synthetic training data generation pipeline (§2) through which we create the Mimir Training Corpus comprising 161 datasets. The pipeline includes synthesis of grammar correction, denoising, and paragraph reordering tasks, and Mimir only receives gradient signal from the completions.
- We provide an extensive evaluation across English, Math & Code, and Danish benchmarks, showing Mimir represents the new state of the art among sub-2B models, and competes with larger frontier models such as Gemma 4 E4B and Qwen 3.5 9B (§4).
- We further report on (a) Mimir's memorisation behavior, including extraction audits of data sources with potential copyright risk (§5); and (b) inference-time ablations that vary both the total recurrent depth and its split between the H- and L-modules (§6).

2 DATA

2.1 SYNTHETIC INSTRUCTION DATA PIPELINE

Satisfying our permissibility criteria rules out important parts of the data used to train HRM-Text 1B (Wang et al., 2026), and permissible Danish instruction data is scarce. We therefore build six complementary pipelines that turn permissible sources into instruction-tuning data. All pipelines share a generate-and-audit design: Gemma 4 31B (Gemma Team, 2026) generates candidate examples and, in a separate call, audits each candidate against pipeline-specific criteria described below. Only accepted examples enter the Mimir training corpus. Further details are in Appendix A.

Turning text corpora into instruction data. Openly licensed text corpora contain rich language but are not in the instruction format Mimir is trained on. We convert texts into supervised tasks, for which the answers originate from the source text itself, i.e., the generator must use text that exists verbatim in the corpus and the auditor checks this. Specifically, we construct four task families: corruption repair (denoising), T5-style span filling, causal-style prefix continuation, and paragraph reordering. We apply these transformations to documents from the Common Pile (Kandpal et al., 2025) for English and from Danish Dynaword (Enevoldsen et al., 2025) for Danish. These instruction data contribute 3.02B and 0.99B tokens per epoch, respectively. The rationale for this design is that it grounds a substantial part of the training signal in human-written text, while the answers remain verifiable against the source. These tasks also suit the PrefixLM attention mask: the corrupted, masked, truncated, or reshuffled text sits in the bidirectionally attended prefix, and the loss is computed only on the original text in the response.

Repairing instead of filtering instruction datasets. Existing instruction datasets often contain flawed answers, including factually incorrect, irrelevant or improperly formatted ones. We audit each conversation and either keep it, regenerate the flawed answer, or discard it if it cannot be repaired; each regenerated answer is re-audited against the original conversation, so that repairs do not drift from the intent of the prompt. For the English OpenHermes (Teknium, 2023) data, this keeps 460,695 of 1,001,551 conversations unchanged (46.0%), replaces 457,400 with accepted repairs (45.7%), and discards 83,456 (8.3%). We record row-level evidence for every decision.

Structure-preserving translation into Danish. To extend high-quality English instruction data to Danish, we machine-translate conversations with the generator model while preserving their structure: roles, code, formulas, labels, and structured outputs remain intact, and only the natural-language content is translated. The rationale is that naive translation often corrupts the elements that make instruction data useful, such as code blocks or answer labels that must match a fixed format. We audit each translation, and whenever the English source is repaired (see above), we retranslate the repaired source and replace the corresponding Danish row, which keeps the two languages aligned. This pipeline yields the Danish OpenHermes data (0.92B tokens per epoch).

Deriving task-specific supervision from domain documents. Agreement-supplied Danish sources, i.e., DBC and Lex.dk, are of high quality but may not be shared publicly. We therefore do not train on the documents directly, but derive instructions from each document's natural structure. For example, book descriptions and reviews yield summarization and characterization tasks, encyclopedia articles yield question answering, and paired documents yield transformation tasks. Because the model only sees derived tasks, this conversion also limits the surface for verbatim memorisation of the source documents (as we tested in our memorization audits in Sec. 5).

Targeting observed data needs. Small models with limited capacity benefit more from data aimed at specific needs than from additional generic instructions (Wang et al., 2025a). In order to target such needs, our synthetic data generation approach is guided by failures observed on intermediate checkpoints. We identify task families and prompt variations relevant for addressing such failures, then generate and audit bilingual examples for each family. This results in six new families: native tool calling, Danish summarization with rewrite controls, multi-turn Danish–English chat, code debugging, strict math answer formats, and constrained format following.

Table 1: Token share and dataset count by functional category.

Category	Tokens/epoch	Share	Datasets	Avg. tokens/dataset
Danish instruction & knowledge	15.56B	22.07%	30	518.6M
English instruction	13.58B	19.26%	16	848.5M
Sapient mixed (Flan/Platypus)	12.00B	17.02%	71	169.0M
Math & reasoning	10.40B	14.76%	10	1,040.2M
Mimir synthetic	7.05B	10.00%	16	440.4M
Agentic & tool use	6.66B	9.46%	8	832.5M
Machine translation	3.50B	4.96%	5	699.0M
Science & summarization	1.74B	2.47%	5	348.5M
Total	**70.48B**	**100%**	**161**	

Converting heterogeneous agent traces into native tool use. Agentic datasets encode tool calls in incompatible formats, often wrapped in legacy conventions that conflict with newer chat templates. We convert these traces into native tool-use supervision by recovering the declared tool schemas, the structured calls, the matching tool results, and the final answers. We remove incompatible legacy wrappers and repair malformed trajectories, for example calls without a matching result. We apply this conversion to Dolci (Team Olmo, 2025), Glaive[1], xLAM (Liu et al., 2024), ToolACE (Liu et al., 2025a), and Nemotron (NVIDIA, 2025) for Mimir v1, and additionally to DeepDive (Lu et al., 2025) for Mimir v1.5.

2.2 Mimir Training Corpus v1

The final training corpus for Mimir v1 covers varying objectives, ranging from English and Danish instruction and knowledge to mathematics and agentic-style post-training data. This mixture, which we denote the Mimir Training Corpus v1, draws on 161 datasets with all but 2 Danish ones being freely available on the Hugging Face Hub. The corpus amounts to **70,479,308,606 tokens** per epoch. Appendix E lists all datasets and their sources.

Language Distribution. Table 7 shows the distribution of tokens by language. The corpus is predominantly English (68.6%), with Danish contributing 24.7%, bilingual Danish–English data a further 6.5%, and translation data between Danish and other languages 0.2%.

Category Distribution. We classify each dataset into one of eight functional categories based on its content and intended use. Table 1 shows that 83% of tokens come from outside the HRM-Text collection, and that the three largest of these categories focus on generative tasks: **English instruction** (13.58B, 19.3%), **Math & reasoning** (10.40B, 14.8%), **Danish instruction & knowledge** (15.56B, 22.1%). These three categories account for 39.54B tokens (56.1% of the corpus), nearly all relying on answer or free-form generation rather than multiple-choice selection. Appendix A provides further details on the Mimir Training Corpus v1.

2.3 Mimir Training Corpus v1.5

The Mimir Training Corpus v1.5 extends v1 to **103,214,604,702** tokens per epoch, 46% more than v1. It replaces instruction datasets flagged by an audit of 100 random samples per dataset with repaired versions. All but the same two Danish datasets are available from the Hugging Face hub.

2.4 Comparison to HRM-Text Data

Notably, the training data for HRM-Text 1B[2] is dominated by multiple-choice classification tasks: Flan, Platypus, and tasksource sub-collections. This is a natural consequence of their source benchmarks: Flan NIV2, Platypus (Reclor, SciBench), and task source (PragmEval, ReClor), which are structured around categorical selection. For Mimir, we shift from multiple-choice toward free-form generation, presenting a substantially harder task for the model to achieve exact match accuracy. This change in data composition better prepares Mimir for open-ended conversation and free-form generation, favoring direct-answer tasks over multiple-choice option selection.

[1]Hugging Face ID: `glaiveai/glaive-function-calling-v2`

[2]Hugging Face ID: `sapientinc/HRM-Text-data-io-cleaned-20260515`

3 MODEL

Architecture. Mimir follows the Hierarchical Reasoning Model Text (HRM-Text) architecture (Wang et al., 2026; 2025b), an instance of the broader class of Looped Transformer models (Giannou et al., 2023; Yang et al., 2024; Saunshi et al., 2025), whereby the same transformer modules are applied repeatedly to iteratively refine the residual stream representation. The architecture is designed in a way that is reminiscent of Kahneman's ideas of thinking, fast and slow (Kahneman, 2011). It consists of two Transformer stacks, both with a PrefixLM attention mask: one L-module for fast, low-level and one H-module for slower, high-level reasoning. To operationalize this idea, the L-module is applied more frequently, in consecutive runs before each H-module application (e.g., three L applications for each of two H applications, denoted as `LLLHLLLH`). The PrefixLM attention mask enables bidirectional attention within the instruction or prompt prefix, while preserving the causal structure for the response to facilitate generation. The states are initialized as $z_H^0 = x$ (the input embedding) and $z_L^{0,0} = z_{L,\text{init}}$ (a learned state). For each high-level cycle i, the L-module is applied N_L times, $z_L^{i,j+1} = f_L(z_L^{i,j} + z_H^i)$ for $j = 0, \ldots, N_L - 1$, after which the H-module updates $z_H^{i+1} = f_H(z_H^i + z_L^{i,N_L})$ and the next cycle continues from $z_L^{i+1,0} = z_L^{i,N_L}$.

The L module takes as input a combination of the previous state of the residual stream and the last-updated H representation. The input embeddings initialize the H-level state, effectively connecting the L module to the input.

Mimir employs a hidden size of 1,536. The model has 12 attention heads per layer and a feed-forward expansion factor of 4. We employ a Rotary Position Embedding (RoPE) (Su et al., 2024) with $\theta = 10{,}000$ and pre-normalization within each block, applying the RMSNorm before both the self-attention and the feed-forward sublayers with $\epsilon = 10^{-6}$. An additional RMSNorm is applied after the final Transformer block before the output head. We provide the full model configuration in Appendix B.1. Mimir has 1.38B non-embedding parameters and 0.40B embedding parameters.

For Mimir v1, backpropagation through the recurrence is truncated to the last five module applications (`HLLLH`); the truncation window grows linearly from two to five applications over the first 20% of optimizer steps. We use the Gemma 4 tokenizer (Gemma Team, 2026), with a 262,144-token vocabulary, and its chat template, which provide efficient multilingual tokenization and native conventions for multi-turn chat and tool calling.

Training Mimir v1. Mimir v1 was trained with Fully Sharded Data Parallelism using bfloat16 as dType for computation and fp32 as gathering precision. We employ the AdamW optimizer (Loshchilov & Hutter, 2019) with a peak learning rate of 3×10^{-4}, 2,000-step linear warm-up, and a constant schedule thereafter. We use a global batch size of 262,144 tokens with a gradient accumulation of 2 on 8 accelerators for a per-accelerator batch size of 16,384, fitting 4 contexts of length 4,096 each. Appendix B.2 provides an overview of training hyperparameters.

We trained Mimir v1 for 1.65M steps on 8 NVIDIA B200 GPUs with 180 GB HBMe3 in just under 3 weeks with an average step time of just under 1.1 seconds, i.e., using approx. 8,000 B200 hours.

Continued Training: Mimir v1.5. We continually trained Mimir for $1.2M$ steps on the Mimir Training Corpus v1.5. For Mimir v1.5, we gradually expand the backpropagation steps from 5 steps for Mimir v1 to full backpropagation, i.e., all 8 steps of the `LLLHLLLH` recurrence.

Because the H- and L-modules reuse their weights across iterations, a single parameter update to a module changes the model's output through every one of its applications. We therefore divide the base learning rate η by the number of applications of each module that receive gradients, i.e., $\eta_M = \eta / k_M$ for $M \in \{H, L\}$, where k_M is determined by the current backpropagation window. For example, with `LLLHLLLH` recurrence and backpropagation truncated to the last five steps (i.e., `HLLLH`), the H-module receives two gradient contributions and the L-module three, which yields $\eta_H = \eta/2$ and $\eta_L = \eta/3$. Specifically, when the truncation window grows from two to five updates during the first 20% of training, the dividers follow the window: at two steps (`LH`), both modules are trained with η, whereas at four steps (`LLLH`), $\eta_L = \eta/3$ and $\eta_H = \eta$. The peak learning rate of 3×10^{-4} refers to η before division. We found that this scaling was critical to stabilize training when increasing the backpropagation to all 8 steps of the recurrence. We close Mimir v1.5 training with a cooldown epoch scaling down the base learning rate to 3×10^{-5}.

Table 2: English benchmark results. Best scores in **bold**.

Band	Model	BoolQ (Acc)	Winogrande (Acc)	Hellaswag (Acc)	MMLU (Acc)	ARC-C (Acc)	DROP (F1)	GovRep. (R1)	*Avg.*
	Mimir v1.5	**91.6**	**81.8**	83.9	67.4	90.4	**88.8**	**40.9**	**77.8**
	Mimir v1	87.8	73.5	67.3	57.5	81.6	83.1	32.0	69.0
	Ouro 1.4B	83.9	56.0	59.6	67.1	87.0	38.9	32.4	60.7
<2B	HRM-Text 1B	87.5	70.4	60.4	58.7	82.2	78.1	25.4	66.1
	Qwen 3.5 0.8B	69.8	48.9	37.0	51.5	68.4	45.2	32.5	50.5
	Gemma 3 1B	62.4	49.1	30.6	37.5	43.5	7.0	29.5	37.1
	OLMo 2 1B	67.2	51.0	42.4	41.6	48.1	12.4	37.7	42.9
	Qwen 3.5 2B	80.8	53.4	64.6	62.8	82.7	31.3	31.5	58.2
2–3B	SmolLM3 3B	84.3	60.3	65.1	60.2	79.5	54.0	38.1	63.1
	Ouro 2.6B	89.0	71.7	78.4	74.5	93.0	52.0	33.1	70.2
	Qwen 3.5 4B	87.0	70.0	83.2	75.8	92.9	48.0	27.9	69.3
4–5B	Gemma 4 E2B	64.1	56.7	55.6	59.3	69.8	57.3	33.6	56.6
	Gemma 4 E2B (thinking)	83.4	63.0	55.8	72.0	86.8	70.8	34.7	66.6
	Qwen 3.5 9B	89.3	75.1	**88.6**	**79.5**	**94.1**	82.1	30.7	77.1
	Apertus 8B	79.7	56.0	66.9	62.1	79.3	52.5	35.0	61.6
8–9B	Ministral 3 8B	88.2	67.2	67.6	73.9	90.1	81.0	31.3	71.3
	Gemma 4 E4B	85.2	64.0	44.6	35.4	49.8	74.0	31.2	54.9
	Gemma 4 E4B (thinking)	88.4	76.1	69.8	68.6	88.8	79.9	34.2	72.3

4 EVALUATIONS

Setup. All benchmarks were evaluated under greedy decoding with shuffle seed 4242 on full test sets. Some English benchmarks use few-shot prompting, with the number of shots following Wang et al. (2026)'s evaluation methodology. All Danish tasks are 0-shot. Further details on the setup and evaluation datasets can be found in Appendix F.

Table 3: Math & Code benchmark results. Best scores in **bold**.

Band	Model	GSM8K (Acc)	MATH (Acc)	HumanEval (Acc)	*Avg.*
	Mimir v1.5	91.7	48.4	73.2	71.1
	Mimir v1	89.9	45.8	56.7	64.1
	Ouro 1.4B	63.5	51.6	54.9	56.6
<2B	HRM-Text 1B	84.8	56.0	0.0	46.9
	Qwen 3.5 0.8B	49.1	36.1	30.5	38.6
	Gemma 3 1B	49.7	37.2	42.7	43.2
	OLMo 2 1B	59.4	18.8	15.9	31.4
	Qwen 3.5 2B	73.7	55.7	47.6	59.0
2–3B	SmolLM3 3B	80.0	62.2	61.6	67.9
	Ouro 2.6B	74.0	59.0	49.4	60.8
	Qwen 3.5 4B	60.5	56.5	78.0	65.0
4–5B	Gemma 4 E2B	88.3	64.2	73.8	75.4
	Gemma 4 E2B (thinking)	90.3	49.1	72.0	70.5
	Qwen 3.5 9B	**95.5**	74.2	**87.8**	**85.8**
	Apertus 8B	66.7	25.4	40.2	44.1
8–9B	Ministral 3 8B	91.7	61.4	75.6	76.2
	Gemma 4 E4B	92.0	71.8	84.8	82.9
	Gemma 4 E4B (thinking)	93.8	**75.3**	77.4	82.2

Reference models. We compare Mimir against 14 models from 0.8B to 9B parameters, including but not limited to Qwen 3.5 0.8B/2B/4B/9B (Qwen Team, 2026), Gemma 4 E2B/E4B, and view HRM-Text 1B, Ouro 1.4B/2.6B (Zhu et al., 2025) as our main baselines.

Results. Tables 2, 3, and 4 show evaluation results across English, Math & Code, and Danish benchmarks, respectively. Mimir v1.5 attains the highest average of all models on English and Danish, and the highest average among models up to 3B on Math & Code.

On English, Mimir v1.5 scores best of all models on BoolQ, Winogrande, DROP, and GovReport, and second to Qwen 3.5 9B on HellaSwag; its average narrowly exceeds that of Qwen 3.5 9B (77.8 vs. 77.1). On Math & Code, it scores best of all models up to 5B on GSM8K and best up to 3B on HumanEval, but trails HRM-Text 1B and Ouro 1.4B on MATH. On Danish, Mimir v1.5 scores best of all models on DaLA, GEC, Daisy, WikiQA, and HellaSwag-DA; Mimir v1 also exceeds every non-Mimir model on DaLA, GEC, and WikiQA, and both versions exceed every other model's Danish average. Relative to Mimir v1, v1.5 improves on every benchmark except Angry Tweets and N.News. Mimir v1.5 outperforms HRM-Text 1B on every benchmark except MATH, and Ouro 2.6B on every benchmark except MMLU, ARC-C, MATH, and PIQA-da.

Table 4: Danish benchmark results. Best scores in **bold**.

Band	Model	Angry Tweets (Acc)	DaLA (F1)	GEC (EM)	PIQA (Acc)	Daisy (EM)	WikiQA (EM)	WMT (chrF)	N.News (chrF)	IFEval (Acc)	Hellaswag-DA (Acc)	*Avg.*
	Mimir v1.5	63.3	**96.8**	**92.5**	73.2	**12.3**	**66.9**	54.6	29.4	66.4	**74.4**	**63.0**
	Mimir v1	67.4	96.1	85.6	53.7	9.6	66.8	53.9	35.9	63.9	35.3	56.8
	Ouro 1.4B	18.8	35.6	0.3	**86.1**	0.5	12.5	27.7	19.0	22.6	25.0	24.8
<2B	HRM-Text 1B	42.4	26.7	0.5	13.0	0.0	34.9	25.4	26.8	18.5	28.8	21.7
	Qwen 3.5 0.8B	53.8	51.0	0.7	56.5	0.7	41.6	37.8	35.3	39.6	25.0	34.2
	Gemma 3 1B	54.4	41.0	3.3	72.2	1.4	42.6	45.1	35.6	47.2	24.8	36.8
	OLMo 2 1B	33.6	48.7	0.2	75.0	0.0	8.4	30.0	33.8	32.5	26.7	28.9
	Qwen 3.5 2B	61.6	36.4	8.0	25.0	2.5	49.4	45.6	34.9	56.1	24.7	34.4
2–3B	SmolLM3 3B	63.2	33.5	3.3	51.9	2.2	0.3	37.3	36.0	49.8	40.1	31.7
	Ouro 2.6B	17.8	33.3	0.8	75.9	1.4	24.8	30.0	19.1	26.8	24.2	25.4
	Qwen 3.5 4B	69.1	50.1	42.6	70.4	4.7	57.1	52.1	**37.0**	73.7	34.7	49.2
4–5B	Gemma 4 E2B	64.6	56.7	36.9	46.3	5.6	44.1	55.2	35.7	75.5	25.6	44.6
	Gemma 4 E2B (think)	67.7	66.8	23.4	63.9	5.1	59.3	56.0	36.3	**81.2**	39.0	49.9
	Qwen 3.5 9B	68.9	70.1	52.0	62.0	8.5	50.4	54.8	22.3	70.8	70.7	53.1
	Apertus 8B	63.5	51.5	37.6	69.4	10.8	59.4	55.8	21.3	56.0	40.0	46.5
8–9B	Ministral 3 8B	63.9	58.9	17.9	63.0	7.3	48.7	50.2	18.4	48.6	45.5	42.2
	Gemma 4 E4B	69.7	63.0	49.4	71.3	7.9	53.5	57.3	22.0	77.3	64.3	53.6
	Gemma 4 E4B (thinking)	**69.9**	72.6	39.7	69.4	7.6	60.6	**57.5**	21.4	80.4	66.2	54.5

5 Memorisation Audits

Training data duplication and multi-epoch training, as conducted for Mimir v1 and v1.5, may generally lead to an increase in memorisation (Kandpal et al., 2022; Lee et al., 2022; Carlini et al., 2023; Slack & Al Moubayed, 2025). Relatedly, regulatory frameworks may distinguish between lawful access, legitimate training, and copyright law. As legitimate training and copyright are two separate matters, studying memorisation behavior is of interest to assess the model's copyright violation risk.

Therefore, we run two memorisation audits on Mimir v1 independently from each other to assess both its propensity and its capability to memorize training data. We cover four data categories: synthetic instruction-tuning data derived from agreement-backed sources (A), instruction-tuning data from Hugging Face sources with uncertain opt-out status (B), instruction-tuning data from Hugging Face sources with high confidence of no applicable opt-out (C), and other low-risk synthetic and reasoning post-training datasets (D). All remaining datasets are either covered by open licenses, in the public domain, or synthetic datasets derived from such openly licensed or public domain datasets.

The first audit with targeted prefix attacks yielded predominantly short matches, verbatim spans of 50 tokens or more were found in only 0.00022%–0.015% of training documents across all categories. Across all model-input evaluations in the second audit, only 0.000045% corresponded to coherent prose and 0.00000073% to expressive prose, with no critical copyright findings, indicating low copyright-related memorisation risk under the tested scenarios. Details are in Appendix C.

6 Test-time Ablations

We study how sensitive the trained model is to changing the cycle pattern at inference time. The question is prompted by the residual-stream dynamics in Appendix D: a similar profile of large writes and rotations recurs with every module application, peaking near module boundaries. If each application performed an interchangeable refinement step, the number of applications could be changed without retraining. We test whether this holds for the L- and H-modules separately.

Setup. We evaluate Mimir v1 under varying recurrence patterns by changing the number of H and L cycles at inference time, keeping the weights fixed. We vary $H \in \{1, 2, 3\}$ and $L \in \{1, \ldots, 5\}$, giving a recurrent depth $D = H(L+1)$ between 4 and 12, and evaluate on three single-token multiple-choice tasks (ARC-C, MMLU, HellaSwag) and two free-generation tasks (GSM8K, GEC-DaLA). Prompting and scoring follow the main evaluation setup (Section 4), but all seven schedules are served by the same vLLM runtime with a different attention backend, so absolute values differ slightly from Tables 2–4.

Table 5: Test-time recurrence sweep. The weights are fixed throughout; only the inference schedule changes. The shaded row is the schedule the model was trained with. Best per column in **bold**. † Collapsed configuration: score below chance or near zero.

D	H	L	Schedule	ARC-C	MMLU	GSM8K	HellaSwag	GEC-DaLA
4	1	3	LLLH	11.09†	6.10†	0.00†	1.61†	0.00†
4	2	1	LHLH	37.29	38.58	46.63	33.30	65.63
6	2	2	LLHLLH	70.82	53.99	81.12	57.96	84.57
8	2	3	LLLHLLLH	79.95	59.13	**86.73**	**66.72**	**85.94**
10	2	4	LLLLHLLLLH	79.69	58.74	84.76	60.94	85.16
12	2	5	LLLLLHLLLLLH	**80.38**	**59.28**	84.38	63.83	80.37
12	3	3	LLLHLLLHLLLH	74.83	52.35	0.53†	57.24	0.68†

Results. Table 5 shows that the trained schedule is best or near-best on every benchmark, so additional recurrent computation at inference does not improve accuracy. Varying the number of L-module applications degrades performance gracefully, whereas varying the number of H-module applications breaks the model, even at matched depth, where only the allocation of recurrent computation differs (D=4: H=2 vs. H=1; D=12: H=2 vs. H=3). A two-sided exact McNemar test on paired per-item GEC-DaLA outcomes (N=1,024) confirms both contrasts, with no discordant pairs in the failing direction (672 vs. 0 and 816 vs. 0; $p < 10^{-200}$). Depth alone therefore does not predict whether the model works, and similar-looking residual dynamics do not imply interchangeable computation. Since we vary a single checkpoint, this shows that the model is specialised to its training-time H, not that H=2 is intrinsically required.

The two failing configurations fail in different ways: H=1 collapses on every task, falling below chance even on multiple choice. H=3 still emits well-formed option letters (at most 0.9% invalid outputs) and loses only 7.1 points on multiple choice on average, yet loses 85.7 points on free generation, failing to produce a parseable answer on 99.3% of GSM8K problems. Evaluating on multiple choice alone would thus understate the damage roughly twelvefold.

7 Related Work

Recurrent Depth. Dehghani et al. (2019) introduced the Universal Transformer, which reuses a shared block of weights across depth to decouple computation from parameter count. Yang et al. (2024) showed that looped transformers match standard transformers on a range of in-context learning problems while using less than 10% of the parameters, and follow-up work has shown the approach's feasibility for language modeling (Geiping et al., 2025). Recent looped language models include Ouro (Zhu et al., 2025), trained on 7.7T tokens at 1.4B and 2.6B parameters, which we use as baselines. Saunshi et al. (2025) found that performance depends on effective depth. Kohli et al. (2026) show that recurrent models succeed at systematic compositional generalization.

Hierarchical reasoning models (Wang et al., 2025b), which introduced the L- and H-module split, were first proposed for algorithmic and puzzle reasoning. HRM-Text (Wang et al., 2026) then showed that such models can be feasible as generalist language models, outperforming scale-matched looped transformers and standard transformers without weight sharing. On these puzzle tasks, the Tiny Recursive Model (Jolicoeur-Martineau, 2025) shows that the hierarchy is not needed: a single 7M-parameter, two-layer network that recursively refines its latent state outperforms HRM.

Synthetic data and transplant approaches. Synthetic datasets have emerged as a strong off-policy distillation strategy. Self-Instruct (Wang et al., 2023) introduced the core generate-and-filter paradigm by bootstrapping a model from its own generations. Alpaca-style pipelines further showed that strong instruction-following behavior can be induced from synthetic teacher outputs (Taori et al., 2023). Magpie (Xu et al., 2025) extended this direction with a pipeline focused on alignment-data synthesis, while also reporting strong instruction-tuning performance with substantially fewer examples than many larger alternatives. Our pipeline builds directly on this line of work: we use staged synthetic generation with quality gates before promoting samples into the training corpus.

A complementary approach grounds synthetic examples in given sources, such as existing datasets or raw text. FineInstructions (Patel et al., 2026) is an example of this direction, scaling synthetic instruction pipelines to pre-training scale. We follow this approach in our transplant strategy, creating datasets with similar capabilities while respecting data rights and replacing faulty datasets when needed.

European efforts on low/medium-resource language modeling. A wide range of initiatives, especially in Europe, are currently building language models with a focus on their local languages, such as Apertus (Apertus et al., 2025), GPT-NL (Van Oort et al., 2026), NorwAI (Gulla et al., 2026) and AI Sweden (Ekgren et al., 2024), as well as broader efforts like OpenEuroLLM (OpenEuroLLM Consortium, 2026). However, most of these efforts operate against the obstacle of collecting and training on massive pretraining corpora for LLM architectures. European data-protection and copyright legislation raises both the cost and complexity of data acquisition and reuse. As a result, many projects train either relatively small models from scratch, limiting their capabilities, or adapt existing foundation models through continued pre- and post-training, which improves performance but does not deliver on full data sovereignty.

Compared to the conceptually most similar models, Mimir v1.5 attains higher domain averages than HRM-Text 1B, Ouro 1.4B, and Ouro 2.6B while training on roughly 10% of Ouro's 7.7T tokens. One plausible contributor is our instruction-formatted training data, although we do not isolate this factor. Mimir v1.5 thus shows that competitive sub-2B language models can be trained from scratch on permissible data only. Models, code, and all data except the two agreement-supplied sources will be made publicly available.

8 Conclusion

In this work, we presented Mimir v1 and v1.5, two sub-2B-parameter language models that leverage the HRM-Text architecture to provide highly competitive performance using only permissible data, excluding data containing personal information or copyright infringement and including data that is either openly licensed, made available by agreement, or allowed by the European Union's text and data mining exception for research institutions. We curated a diverse corpus, Mimir Training Corpus, consisting of 70.5B (v1) and more than 100B tokens per epoch (v1.5), including synthetic instruction datasets. We have shown that strong instruction-following and reasoning capabilities can be achieved without web-scale pre-training corpora or non-permissible data sources. Mimir improves substantially over HRM-Text and Ouro, and is competitive with, or superior to, much larger models on several English and Danish benchmarks. It still lags behind Gemma 4 (5B, effective 2.3B) on the Math & Code domains, making room for improvement in future iterations.

Limitations While Mimir demonstrated strong performance across all domains, punching above its weight class, several limitations apply. First, its capabilities are shaped by the domains represented in the training mixture, which covers primarily English and Danish language tasks, mathematics, coding, and factual question answering. Because Mimir is not pretrained on broad web corpora, its performance may be limited on domains, styles, or factual areas that are weakly represented or absent from the mixture, more so than for conventionally pretrained models, which acquire broad world knowledge through web-scale next-token prediction. Mimir is not multilingual, and while its capabilities may generalize to understand other languages, it is a bilingual model trained with focus on English and Danish. A further limitation concerns the amount and type of information that can be encoded in the relatively small model parameters. Even though Mimir exhibits good performance, factual knowledge is limited. Last, Mimir's inference time measurements scale, as expected based on effective depth at inference time, making it achieve throughput similar to a conventional non-looped LLM with an equivalent number of layers.

Future work will focus on investigating scaling behavior of HRM models like Mimir and reinforcement learning, which is yet unexplored for this architecture. Lastly, we will continue developing the dataset to test the limits of synthetic instruction data generation.

ETHICS STATEMENT

Training memory-efficient architectures contributes to the democratization of AI research, enabling practitioners to inference on consumer hardware. In principle, open-source models come with risks of misuse. However, we believe that those risks are marginal given the small size of our model. To quantify copyright infringement risks, we ran two memorisation audits independently from each other, and found that only a negligible fraction of training data can be reproduced. Upon qualitative inspection, many of those were mathematical formulas and a popular children's song.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, implement methods and experiments, clean and reformat datasets, and provide feedback on research methodology or experiments (mock review). We have not used generative AI tools to interpret results. Development of theoretical models or conceptual frameworks, formulation of mathematical claims or critical ingredients for proving mathematical claims, proofs writing and hypotheses proposal or refinement are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated evaluation code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Both models, Mimir v1 and Mimir v1.5, the collection of synthetic training datasets, and the training and evaluation code, will be made publicly available. Evaluations were implemented and run in InspectAI (AI Security Institute, 2024) and are based on publicly available benchmark datasets. Details of the evaluation methodology are provided in Appendix F. Architecture and training hyperparameters are listed in Appendix B.

REFERENCES

UK AI Security Institute. Inspect AI: Framework for Large Language Model Evaluations, 2024. URL https://github.com/UKGovernmentBEIS/inspect_ai.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. Apertus: Democratizing open and compliant LLMs for global language environments, 2025. URL https://arxiv.org/abs/2509.14233.

Gianluca Barmina, Nathalie Carmen Hau Norman, Peter Schneider-Kamp, and Lukas Galke Poech. DaLA: Danish linguistic acceptability evaluation guided by real world errors. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pp. 4312–4326, 2026a.

Gianluca Barmina, Peter Schneider-Kamp, and Lukas Galke Poech. LLMs can leak training data but do they want to? a propensity-aware evaluation of memorization in LLMs, 2026b. URL https://arxiv.org/abs/2606.06286.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akriti Kuri, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Carolina Condez, Anamaria-Roberta Hartl, Andiswa Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anjali Chaudhary, Anna Vacalopoulou, Annette Kuuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Anupam Nayak, Ashvanth S, Ayodeji Lana, Ayu Purwarianti, Bashar Alhafni, Benedict Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska urić, Blessing Ogundipe, Bolaotan Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Çınar, Cade Kane, Can Udomcharoenchaikit, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Nwokolo, Christopher Homan, Clément Sampebgo, Cristina España-Bonet, Cynthia Amol, Daeyoep Lee, Dan Saattrup Smart, Dana Arad, Daniil Dzenhaliou, Dasol Choi, David Liu, David Semedo, David Anugraha, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dennis Owusu, Dhyuthy Krishna Kumar, Diogo Tavares, Diogo Glória-Silva, Divyanshu Goyal, DongGeon Lee, E. Kelly Buchanan, Ebele Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elias Herranen, Eliza Acharya, Eman Nisar, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Itiola, Faran Taimoor Butt, Fareeha Fayyaz Sheikh, Fathima Thekkekara, Fatima Haouari, Faustin Nsengiyumva, Fenal Ashokbhai Ilasariya, Filbert Aurelian Tjiaranata, Firas Laakom, Francesca Grasso, Francesco Periti, Francesco Orabona, Gbenga Kayode Solomon, Genta Indra Winata, Gia Nghia Ngo, Gloria Udhedhe-oze, Gonçalo Vinagre, Gopi Naga Sai Ram Challagolla, Gorka Urbizu-Garmendia, Gouthami Vadithya, Guijin Son, Gulnaz Abdykadyrova, Gyan Swaroop Mohapatra, Hafeez Ullah, Hafsteinn Einarsson, Hai Hu, Hamidreza Saffari, Hamza Zaidi, Haopeng Zhang, Harethah Abu Shairah, Harry Vuong, Hele-Andra Kuulmets, Hitesh Laxmichand Patel, Houda Bouamor, Hwanjo Yu, Iben Nyholm Debess, İbrahim Ethem Deveci, Ikhlasul Akmal Hanif, Ikhyun Cho, Inês Vieira, Inês Calvo, Isaac Manzi, Ismael Illa Salifou, Ismail Daud, Ismail Yusuf, Itay Itzhak, Ivan Zhelyazkov, Ivan Belashkin, Ivan Spada, Jacob Brinton, Jafar Isbarov, Jaka Čibej, Jan Kocoń, Jan Cuhel, Jauza Krito, Jebish Purbey, Jennifer Za, Jennifer Mickel, Jenny Kunz, Jessica Ratovondranto, Jeyarajalingam Varsha, Jihae Jeong, Jimena Tena Dávalos, Jinu Lee, João Magalhães, John Seon Keun Yi, Jongin Kim, Joseph Chataignon, Joseph Marvin Imperial, Jubeerathan Thevakumar, Judith Land, Julia Alekseenko, Junchen Jiang, Jungwhan Kim, Kairit Sirts, Kamesh R, Kamesh V, Kanda Tshinu, Kätriin Kukk, Kaustubh Ponkshe, Kavsar Huseynova, Ke He, Kenneth Enevoldsen, Kent Joshua Alvarez, Kerem Zaman, Khalil Mrini, Kian Kyars, Komal Gour, Krishnakumar Lainitha, Krister Kruusmaa, Kunal Mukherjee, Kusum Chouhan, Laura Castro, Laura M. Porrino-Moscoso, Lenny Sivi Za Nzambi, Leshem Choshen, Levent Sencan, Lilja Øvrelid, Lisa Alazraki, Loretta Oma Jones, Lovina Ehimen-Ugbede, Luheerathan Thevakumar, Luxshan Thavarasa, Mahnoor Malik, Mamadou K. Keita, Mansi Jangid, Marco De Santis, Marcos Garcia, Marek Šuppa, Mariam D'Ciofalo, Marii Ojastu, Marium Attaullah, Maryam Sikander, Mausami Narayan, Maximos Skandalis, Mehak Mehak, Mehmet İlteriş Bozkurt, Melaku Bayu, Menan Velayuthan, Mhasilenuo Vizo, Michael Leventhal, Michał Marcińczuk, Mina Almasi, Mirna Potočnjak, Mithil Bangera, Mohammadamin Shafiei, Mohiba Ansari, Mridul Sharma, Mrityunjaya Indoria, Mughees Ur Rehman, Muhammad Ravi Shulthan Habibi, Murat Kolić, Murat Barkın Kınay, Nada Galant, Naina Singh Rathore, Naphat Permpredanun, Narada Maugin, Nathalie Norman, Nicholas Kluge Corrêa, Nikola Ljubešić, Nirmal Thomas, Nisansa de Silva, Nisheeth Joshi, Nitish Ponkshe, Nizar Habash, Nneoma Udeze, Noel Thomas, Noémi Ligeti-Nagy, Nouhoum Coulibaly, Odunayo Ogundepo, Odunayo Kareemat Buliaminu, Oghojafor Godswill Fejiro, Okechukwu God'spraise, Olanrewaju Samuel, Olaoye Deborah Oluwaseun,

Olasoji Akindejoye, Olga Snissarenko, Onyinye Anulika Chiemezie, Orkun Kınay, Osman Tursun, Oyelade Oluwafemi Joshua, Oyesanmi Fiyinfoluwa, Pablo Rodríguez, Pablo Gamallo, Palak Arora, Pedro Valente, Peter Rupnik, Philip Oghenesuowho Ekiugbo, Prakhar Agarwal, Pramit Sahoo, Prokopis Prokopidis, Pua Niau-Puhipau, Quadri Yahya, Rachele Mignone, Raghav Singhal, Rahul Raja, Ram Mohan Rao Kadiyala, Raphael Merx, Rasmus Larsen, Ratnavel Rajalakshmi, Rishav Ghosh, Romina Oji, Ron Kekeha Solis, Rui Guerra, Rushikesh Zawar, Sa'ad Nasir Bashir, Saeed Alzaabi, Sahil Sandeep, Sai Pavan Batchu, Sai Sandeep Kantareddy, Saleha Muzammil, Salsabila Zahirah Pranida, Sam Buchanan, Samuel Rutunda, Sander Land, Sarah Sulollari, Sardar Ali, Saroj Sapkota, Sarveswaran Kengatharaiyer, Saulius Tautvaisas, Sayambhu Sen, Sayantani Banerjee, Sebastien Diarra, Segun Afolayan, Senthilnathan M, Sewoong Lee, Shaan Shah, Shankar Venkitachalam, Sharifa Djurabaeva, Sharon Ibejih, Shivanya Shomir Dutta, Siddhant Gupta, Silvia Paniagua Suárez, Sina Ahmadi, Sivasuthan Sukumar, Siyuan Song, Snegha A, Sokratis Sofianopoulos, Sona Elza Simon, Sonja Benčina, Sophie Gvasalia, Sphurti More, Spyros Dragazis, Stefan Milosavljević, Stephan P. Kaufhold, Suba S, Sultan Alrashed, Surangika Ranathunga, Taiga Someya, Taja Kuzman Pungeršek, Tal Haklay, Tasi'u Jibril, Tatsuya Aoyama, Tea Abashidze, Terenz Jomar Dela Cruz, Terra Blevins, Themistoklis Nikas, Theresa Idoko, Thu Mai Do, Tilek Chubakov, Tina Munda, Tobiloba Owoeye, Tommaso Gargiani, Uma Rathore, Uni Johannesen, Uwuma Ugwu, Vallerie Alexandra Putra, Vanya Bannihatti Kumar, Varvara Arzt, Vasily Konovalov, Vasudevan Nedumpozhimana, Viktoria Ondrejova, Viktoryia Horbik, Vishnu Vardhan Reddy Kummitha, Vuk Dinić, Walelign Sewunetie, Winston Wu, Xiaojing Zhao, Yacouba Diarra, Yaniv Nikankin, Yash Mathur, Yash Bagla, Yeshil Bangera, Yixi Chen, Yiyuan Li, Yolanda Xavier, Yonatan Belinkov, Zaid Alyafeai, Zhargal Batozargalova, Zhengyang Shan, Zhi Rui Tam, Zilu Tang, Zuzana Nadova, Baber Abbasi, Stella Biderman, David Stap, Duygu Ataman, Fabian Schmidt, Hila Gonen, Jiayi Wang, and David Ifeoluwa Adelani. Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures, 2026. URL https://arxiv.org/abs/2510.24081.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint 2107.03374*, 2021.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, jun 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the Ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers, 2019. URL https://arxiv.org/abs/1807.03819.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, et al. Wmt24++: Expanding the

language coverage of wmt24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12257–12284, 2025.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, 2019. URL https://arxiv.org/abs/1903.00161.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. Gpt-sw3: An autoregressive language model for the scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7886–7900, 2024.

Kenneth Enevoldsen, Kristian Nørgaard Jensen, Jan Kostkan, Balázs Szabó, Márton Kardos, Kirsten Vad, Johan Heinsen, Andrea Blasi Núñez, Gianluca Barmina, Jacob Nielsen, Rasmus Larsen, Peter Vahlstrup, Per Møldrup Dalum, Desmond Elliott, Lukas Galke, Peter Schneider-Kamp, and Kristoffer Nielbo. Dynaword: From one-shot to continuously developed datasets, 2025.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38: 41340–41391, 2025.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pp. 11398–11442. PMLR, 2023.

Jon Atle Gulla, Peng Liu, and Lemei Zhang. NorwAI's large language models: Technical report. *arXiv preprint arXiv:2601.03034*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization, 2021.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*, 2025.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, John Kirchenbauer, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben Allal, Elie Bakouch, John David Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, Tom Goldstein, Brian R. Bartoldson, Bhavya Kailkhura, and Tyler Murray. The common pile v0.1: An 8tb dataset of public domain and openly licensed text, 2025. URL https://arxiv.org/abs/2506.05209.

Oliver Kinch. Nordjylland news summarization. Hugging Face, 2023. URL https://huggingface.co/datasets/alexandrainst/nordjylland-news-summarization.

Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers. *arXiv preprint arXiv:2604.07822*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Rasmus Larsen, Nathalie Hau Sørensen, and Kenneth Enevoldsen. IFEval-da. Hugging Face, 2025. URL https://huggingface.co/datasets/danish-foundation-models/ifeval-da.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

Weiwen Liu, Xu Huang, Xingshan Zeng, Xinlong Hao, Shuai Yu, et al. ToolACE: Winning the points of LLM function calling. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2409.00920.

Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. *arXiv preprint arXiv:2506.13284*, 2025b.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, et al. APIGen: Automated pipeline for generating verifiable and diverse function-calling datasets. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.18518.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. DeepDive: Advancing deep search agents with knowledge graphs and multi-turn RL. *arXiv preprint arXiv:2509.10446*, 2025.

Dan Nielsen. ScandEval: A benchmark for Scandinavian natural language processing. In Tanel Alumäe and Mark Fishel (eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 185–201, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.20/.

Jacob Nielsen, Stine L. Beltoft, Peter Schneider-Kamp, and Lukas Galke Poech. SDUs DAISY: A benchmark for danish culture, 2026. URL https://arxiv.org/abs/2601.19930.

NVIDIA. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning. *arXiv preprint arXiv:2512.20848*, 2025.

OpenEuroLLM Consortium. Deriving scaling laws for openeurollm models: Learning rate, batch size and loss, 2026.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. Fineinstructions: Scaling synthetic instructions to pre-training scale, 2026. URL https://arxiv.org/abs/2601.22146.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. Danlp: An open-source toolkit for danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 460–466, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-09-25.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 14855–14881, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2676109d49d1eb26d6bc584a8f556305-Paper-Conference.pdf.

Dean L. Slack and Noura Al Moubayed. Early detection and reduction of memorization for domain adaptation and instruction tuning. *Transactions of the Association for Computational Linguistics*, 13:1459–1473, 2025. doi: 10.1162/tacl.a.49. URL https://aclanthology.org/2025.tacl-1.66/.

Dan Saattrup Smart. Multiwikiqa: A reading comprehension benchmark in 300+ languages, 2026. URL https://arxiv.org/abs/2509.04111.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Team Olmo. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

Jesse J. Van Oort, Frank Brinkkemper, Erik de Graaf, Bram Vanroy, and Saskia Lensink. GPT-NL public corpus: A permissively licensed, Dutch-first dataset for LLM pre-training. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 6893–6903, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/5fbtc336wwx2. URL https://aclanthology.org/2026.lrec-1.548/.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *ACM Trans. Intell. Syst. Technol.*, 16(6), November 2025a. ISSN 2157-6904. doi: 10.1145/3768165. URL https://doi.org/10.1145/3768165.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025b. URL https://arxiv.org/abs/2506.21734.

Guan Wang, Changling Liu, Chenyu Wang, Cai Zhou, Yuhao Sun, Yifei Wu, Shuai Zhen, Luca Scimeca, and Yasin Abbasi Yadkori. Hrm-text: Efficient pretraining beyond scaling. *arXiv preprint arXiv:2605.20613*, 2026.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *International Conference on Learning Representations*, volume 2025, pp. 76346–76382, 2025.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International conference on learning representations*, volume 2024, pp. 42195–42214, 2024.

Ted Zadouri, Markus Hoehnerbach, Jay Shah, Vijay Thakkar, and Tri Dao. Flashattention-4: Algorithm and kernel pipelining co-design for asymmetric hardware scaling. *Proceedings of Machine Learning and Systems*, 8:912–926, 2026.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

A Details of the Mimir Training Corpus

Table 6: Token share by data form.

	Reformatted	Curated + reformatted	Synthetic + audited	Tool-call formatted	Translated + audited	Agreement-supplied	Derived task
Tokens/epoch	46.49B	11.92B	7.81B	1.87B	1.59B	0.67B	0.13B
Share	65.96%	16.91%	11.08%	2.65%	2.26%	0.95%	0.18%

Table 7: Token share by language.

Language	**English (en)**	**Danish (da)**	**Bilingual da+en**	**Other**
Tokens/epoch	48.36B	17.44B	4.61B	0.14B
Share	68.62%	24.74%	6.54%	0.20%

A.1 Data Processing and Harmonization

Datasets enter the corpus in seven distinct forms, reflecting the pipeline's processing stages. We report the forms in Table 6. **Reformatted** datasets are existing Hugging Face repositories simply converted into the training format – this represents the default pathway for public data. **Curated + reformatted** applies to the Sapient mega-repository, whose 107 sub-collections were selected as a curated subset before reformatting. **Synthetic + audited** data is LLM-generated using Gemma4 31B and quality-audited before inclusion, with acceptance rates ranging from single digit percentages to high nineties for different categories. This includes transformations from high-quality English and Danish text corpora into span-filling, denoising, reordering, and continuation tasks. Here, we employ Common Pile (Kandpal et al., 2025) for English text corpora and Danish Dynaword (Enevoldsen et al., 2025) for Danish ones. We similarly generated and audited the `mtc1-synthetic-*` instruction datasets and Sapient-synth transplants. **Tool-call formatted** data incorporates native tool-calling structure for agentic training. **Translated + audited** covers the OpenHermes-based data and DA/EN translations that were repaired and audited. A small corpus of **agreement-supplied** data comes from agreements, where licensing does not permit public sharing. Note that we use such datasets to derive synthetic instruction-tuning data rather than train directly on the data. **Derived task** data is derived from an existing dataset to create a new task formulation.

Table 8: Top 10 datasets by sampled tokens per epoch.

Dataset	Tokens/epoch	Share
sapientinc/HRM-Text-data-io-cleaned-20260515	11.92B	16.91%
danish-foundation-models/laerebogen	8.32B	11.81%
nvidia/OpenMathInstruct-2	6.60B	9.37%
nvidia/Nemotron-SFT-Agentic-v2	4.27B	6.06%
danish-foundation-models/dfm-dyna-instruct	3.54B	5.03%
allenai/Dolci-Instruct-SFT-No-Tools	3.49B	4.95%
ANONYMIZED/opus-da-en-permissive	2.90B	4.12%
allenai/Dolci-Instruct-SFT	2.24B	3.17%
nvidia/AceReason-1.1-SFT	1.95B	2.76%
allenai/big-reasoning-traces	1.66B	2.35%

A.2 Details of Task Category Distribution

The three largest categories: Danish instruction & knowledge (22.07%), English instruction (19.26%), and selected datasets from the Sapient mixed collection (17.02%)[3], together account for over 58% of the corpus. Math & reasoning contributes a further 14.8%, bringing the combined share of the top four categories to 73%. The Danish instruction & knowledge category is the largest by token volume, driven primarily by `lærebogen`, a Danish instruction-following dataset contributing 8.32B tokens (11.8% of the corpus) at 4× repetition, alongside `dfm-dyna-instruct` (3.54B) and `synquid_wiki-instruct-da` (0.99B). English instruction is anchored in Dolci (Team Olmo, 2025) (7.71B combined), Tulu 3 (Lambert et al., 2024) variants (1.57B), and Nemotron instruction-following data (1.60B). The Sapient mixed category is dominated by a single large repository (11.92B, 16.9%), which bundles 107 sub-collections from Flan, Platypus, and tasksource. An additional 70 *Sapient-synth* transplant datasets contribute 75M tokens. These are synthetic recreations of English instruction tasks (Flan NIV2, Flan Dialog, Platypus, Tasksource) in generated-and-audited form, replacing original Sapient data that was non-compliant with our permissible data philosophy. Math & reasoning is led by `OpenMathInstruct-2` (Toshniwal et al., 2024) (6.60B, 9.4%), the third-largest single dataset overall, followed by AceReason-1.1-SFT (Liu et al., 2025b) (1.95B) and verifiable reasoning traces (0.68B). The remaining four categories – synthetic, agentic & tool use, machine translation, and science & summarization – together contribute 19.0B tokens (27%).

A.3 Data Concentration and Repetition

Table 8 shows the top 10 datasets by sampled tokens per epoch. The corpus is highly concentrated: these ten datasets account for 66.5% of all tokens, and the top three alone for 38.1%, while the remaining 151 datasets contribute 33.5%. Two single sources exceed 10% each: the Sapient mega-repository (16.9%) and lærebogen (11.8%). The Dolci family contributes 5.73B tokens across two instruction datasets, the largest English instruction contribution after the cleaned Sapient corpus, and NVIDIA sources account for 12.82B tokens across the reasoning, math, and agentic datasets in the top 10. This concentration is partly by construction: Several datasets are sampled more than once per epoch. Lærebogen is repeated $4\times$, inflating 2.08B base tokens to 8.32B and making it the second-largest entry in the corpus. Eight small Danish datasets are repeated 10x to ensure sufficient coverage despite limited source data, and Dolci-Instruct-SFT-No-Tools is doubled to increase its English instruction contribution. The most heavily repeated dataset, kaenguruen ($20\times$), is negligible in volume at 638K tokens.

B Hyperparameters

B.1 Model Hyperparameters

Table 9 lists the hyperparameters of the model.

[3] `sapientinc/HRM-Text-data-io-cleaned-20260515`

Table 9: Model hyperparameters.

Hidden size	Layers	Half layers	Attn. heads	Expansion	H cycles	L cycles	BP max steps	BP warmup ratio	Pos. emb.	RoPE θ	Norm	Norm ϵ
1,536	32	true	12	4	2	3	5	0.2	RoPE	10,000	pre-norm	10^{-6}

Table 10: Training hyperparameters.

Base LR	LR warmup	LR min ratio	Adam β_1/β_2	Weight decay	EMA decay	Global batch	Grad. accum. steps	Seed
3×10^{-4}	2,000	1.0	0.9 / 0.95	0.1	0.9999	262,144	2	0

B.2 Training Hyperparameters

Table 10 lists the hyperparameters of Mimir v1 training. Continued training for Mimir v1.5 kept all parameters unchanged but added learning rate scaling (as described in the main part), and finished with a cool-down epoch to 10×10^{-5}.

C Memorisation Audit Details

We run two audits independently from each other across four data categories: (A) synthetic instruction-tuning data generated from seeds originating from sources covered by agreements; (B) instruction-tuning data from Hugging Face sources for which we could not reliably determine for every row whether an applicable opt-out had been exercised; (C) instruction-tuning data derived from Huggingface sources for which we have high confidence that no applicable opt-out had been exercised for any of the rows; and (D) other synthetic and reasoning data for which we identified no suspicion of material copyright-related concern but where no coherent open or licensing or public domain status could be ascertained.

First Audit

The first follows the memorisation and propensity methodology of Barmina et al. (2026b), evaluating 1,000 ordinary non-adversarial prompts (called *generic*) across Danish and English, together with 500 targeted extraction prompts per source category, formed using 50-token source-specific prefixes, for a total of 3,000 prompts. The analysis covers 140,990,504 training documents overall. For each generated output, the method checks against every possible suffix of every document in the analysed collection. Figure 2 shows the exact-match length distribution for all settings. Across the generic settings, the verbatim matches reaching 50 or more tokens are rare: they occur only once in Generic EN/B (0.000056%) and once in Generic EN/D (0.0000029%), with none observed in the other English comparisons or any Danish comparison. Under prefix attacks, verbatim text of any length is identified in 0.073% of A documents (8,731/11,989,549), 0.887% of B (15,921/1,794,829), 0.029% of C (26,642/92,415,358), and 0.046% of D (16,072/34,790,768). Detected matches are predominantly short: verbatim spans of 50 tokens or more were found in only 26 (0.00022%), 271 (0.015%), 293 (0.00031%), and 257 (0.00074%) of all checked training documents across A–D, respectively, and average longest spans are 8.8, 17.1, 12.4, and 22.3 tokens. Inspection of the matched material shows that A and C include some short natural-language content, while many of the longer matches in B and D consist of numerical sequences, formulas, matrices, code, algorithms, structured tasks, and mathematical problem solving.

Second Audit

The second audit independently focuses on the type and copyright relevance of reproduced content. Under a fixed greedy 64+64-token protocol, it identified 5,562 exact-match occurrences across 136,612,444 model-input evaluations (0.0041%), corresponding to 3,423 unique source-prefix–continuation pairs. By category, this comprised 15/2,732,080 matches in A (0.0005%), 7/411,508 in B (0.0017%), 4,874/124,246,748 in C (0.0039%), and 666/9,222,108 in D (0.0072%). These were adjudicated by an LLM and subsequently reviewed by a human for coherence, expressiveness, repetition, and potential copyright significance. Only 61/5,562 (1.10%) contained coherent prose and one (0.018%) was classified as expressive prose. Thus, the total potentially risky generated outputs

with respect to all model-input evaluations are 62/136,612,444 (0.000045%). No high-priority copyright findings were identified, and the single medium-priority case was a predictable continuation of a traditional repetitive song. The remaining prose-like matches were primarily factual captions, news sentences, or text repeated locally within the sources.

RESULTS

Taken together, these analyses show no evidence of systematic reconstruction of long, distinctive expressive passages and indicate a generally low copyright-related memorisation risk. These results can be assumed to be indicative of a genuine low level of memorisation. They do, however, not fully exclude that extraction under adaptive prompts, longer prefixes, or alternative attack or decoding strategies might be more successful.

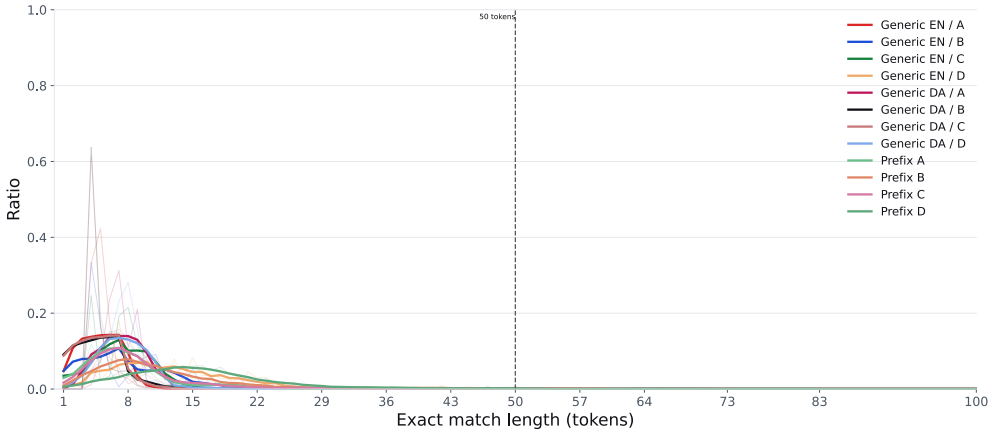


Figure 2: Distribution of verbatim matches for each adversarial (A/B/C/D) and non-adversarial (Generic EN, DA) setting. Each non-adversarial setting is applied separately to every dataset. Solid lines are moving averages, faded lines are the raw data.

Table 11: Summary of the first memorisation audit. For generic settings, matched documents are reported separately for each language–dataset comparison. The match rate is the share of training documents with any detected verbatim match; the final rate reports the share of all checked training documents containing a verbatim span of at least 50 tokens.

Setting	Training documents	Matched documents	Match rate	Matches ≥50 tokens	≥50-token rate	Avg. longest span
Generic EN / A	11,989,549	253	0.0021%	0	0.0%	–
Generic EN / B	1,794,829	339	0.0189%	1	0.000056%	–
Generic EN / C	92,415,358	577	0.0006%	0	0.0%	–
Generic EN / D	34,790,768	464	0.0013%	1	0.0000029%	–
Generic DA / A	11,989,549	1,977	0.0165%	0	0.0%	–
Generic DA / B	1,794,829	1,710	0.0953%	0	0.0%	–
Generic DA / C	92,415,358	2,322	0.0025%	0	0.0%	–
Generic DA / D	34,790,768	1,993	0.0057%	0	0.0%	–
Prefix A	11,989,549	8,731	0.073%	26	0.00022%	8.8 tokens
Prefix B	1,794,829	15,921	0.887%	271	0.015%	17.1 tokens
Prefix C	92,415,358	26,642	0.029%	293	0.00032%	12.4 tokens
Prefix D	34,790,768	16,072	0.046%	257	0.00074%	22.3 tokens

D RESIDUAL STREAM DYNAMICS

Figure 3 shows the residual stream dynamics per dataset measured via relative update magnitude of the residual stream and rotation via centered cosine similarity.

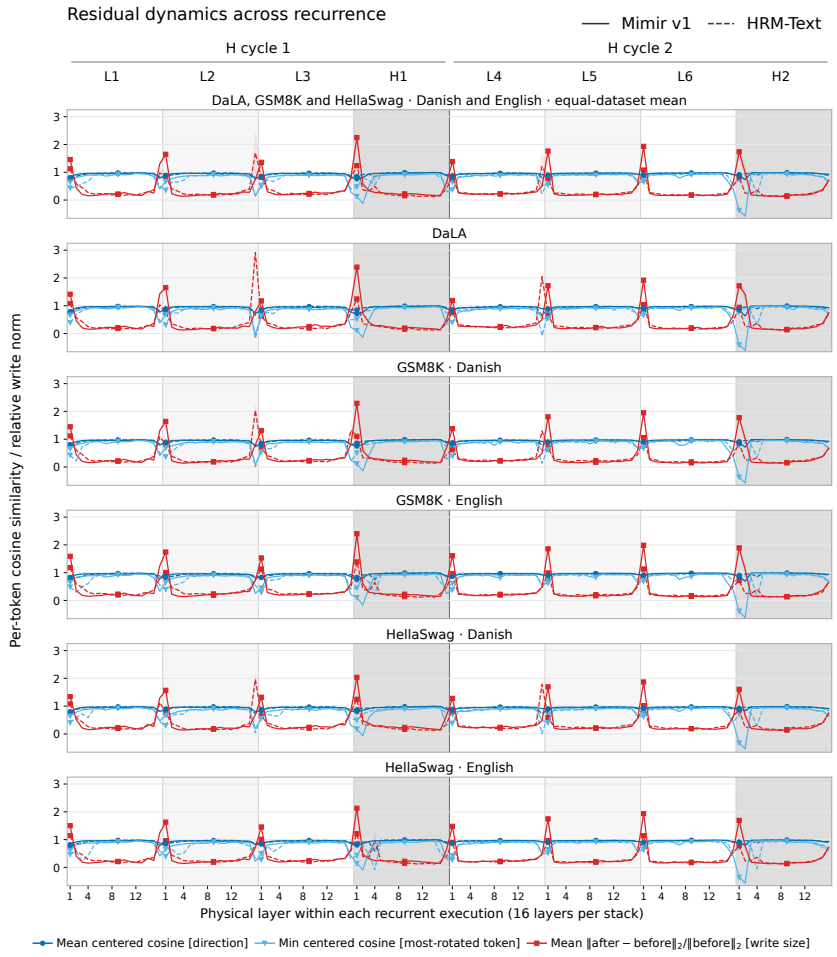


Figure 3: Residual-stream dynamics of Mimir v1 and HRM-Text across DaLA, GSM8K-Danish, GSM8K-English, HellaSwag-Danish and HellaSwag-English. The top panel shows the equally weighted dataset mean; subsequent panels show individual datasets. Measurements include prompt tokens and up to 32 greedily generated continuation tokens. At each physical-layer execution, we measure direction preservation (mean centered cosine similarity), the most-rotated token (minimum centered cosine similarity per sequence, averaged across sequences), and relative write size (the norm of the update divided by the pre-update norm). Both models exhibit recurring peaks in write size and rotation near module boundaries, with broadly consistent patterns across datasets but differences in peak magnitude; lower cosine similarity indicates greater rotation.

Table 12: Summary of the second memorisation audit based on exact 64-token continuations and subsequent content-level adjudication. The 5,562 exact-match occurrences correspond to 3,423 unique source-prefix–continuation pairs.

Category	Model-input evaluations	Exact 64-token matches	Match rate	Coherent prose	Expressive prose
A	2,732,080	15	0.0005%	0	0
B	411,508	7	0.0017%	0	0
C	124,246,748	4,874	0.0039%	40	0
D	9,222,108	666	0.0072%	21	1
Total	**136,612,444**	**5,562**	**0.0041%**	**61 (1.10%)**	**1 (0.018%)**

E List of Training Datasets for Mimir v1

Table 13 lists all the datasets with their corresponding HuggingFace identifier, the form type, how many tokens they represent, and the share in percentages.

Table 13: All 161 datasets in the Mimir corpus, sorted by sampled tokens per epoch.

#	Dataset	Form	Tokens/epoch	Share
1.	sapientinc/HRM-Text-data-io-cleaned-20260515	Curated + reformatted	11.92B	16.91%
2.	danish-foundation-models/laerebogen	Reformatted	8.32B	11.81%
3.	nvidia/OpenMathInstruct-2	Reformatted	6.60B	9.37%
4.	nvidia/Nemotron-SFT-Agentic-v2	Reformatted	4.27B	6.06%
5.	danish-foundation-models/dfm-dyna-instruct	Reformatted	3.54B	5.03%
6.	allenai/Dolci-Instruct-SFT-No-Tools	Reformatted	3.49B	4.95%
7.	[ANONYMIZED]/opus-da-en-permissive	Reformatted	2.90B	4.12%
8.	allenai/Dolci-Instruct-SFT	Reformatted	2.24B	3.17%
9.	nvidia/AceReason-1.1-SFT	Reformatted	1.95B	2.76%
10.	allenai/big-reasoning-traces	Reformatted	1.66B	2.35%
11.	allenai/Dolci-Instruct-SFT-Tool-Use	Tool-call formatted	1.61B	2.29%
12.	nvidia/Nemotron-SFT-Instruction-Following-Chat-v2	Reformatted	1.60B	2.27%
13.	allenai/tulu-3-sft-mixture	Reformatted	1.57B	2.23%
14.	laion/Scientific-Summaries	Reformatted	1.27B	1.80%
15.	[ANONYMIZED]/common-pile-prefix-continuation	Synthetic + audited	1.15B	1.64%
16.	synquid/wiki-instruct-da	Reformatted	988M	1.40%
17.	[ANONYMIZED]/mtc1-openhermes-da	Translated + audited	922M	1.31%
18.	[ANONYMIZED]/common-pile-denoising	Synthetic + audited	884M	1.25%
19.	allenai/tulu-v2-sft-mixture	Reformatted	840M	1.19%
20.	[ANONYMIZED]/common-pile-span-filling	Synthetic + audited	822M	1.17%
21.	[ANONYMIZED]/mtc1-openhermes-en	Translated + audited	672M	0.95366%
22.	allenai/tulu-v2-sft-long-mixture	Reformatted	605M	0.85902%
23.	allenai/verifiable-reasoning-filtered-gpt-41	Reformatted	605M	0.85798%
24.	open-thoughts/OpenThoughts2-1M	Reformatted	534M	0.75834%
25.	[ANONYMIZED]/mtc1-synthetic-native-tool-calling	Synthetic + audited	513M	0.728%

continued from previous page

#	Dataset	Form	Tokens/epoch	Share
26.	[ANONYMIZED]/transformations-danish-danish	Synthetic + audited	496M	0.70403%
27.	[ANONYMIZED]/transformations-english-english	Synthetic + audited	471M	0.66842%
28.	[ANONYMIZED]/transformations-danish-english	Synthetic + audited	469M	0.66529%
29.	[ANONYMIZED]/transformations-english-danish	Synthetic + audited	418M	0.59274%
30.	nvidia/Nemotron-SFT-Multilingual-v1	Reformatted	413M	0.58565%
31.	[ANONYMIZED]/mtc1-synthetic-danish-summarization-rewrite-controls	Synthetic + audited	396M	0.56181%
32.	MegaScience/TextbookReasoning	Reformatted	374M	0.5311%
33.	[ANONYMIZED]/mtc1-synthetic-multiturn-danish-english-chat	Synthetic + audited	366M	0.51997%
34.	DBC (agreement-supplied)	Agreement-supplied	356M	0.50529%
35.	[ANONYMIZED]/danish-dynaword-denoising	Synthetic + audited	323M	0.45861%
36.	Lex.dk articles	Agreement-supplied	313M	0.44467%
37.	oliverkinch/machine-translation-da-en	Reformatted	284M	0.40348%
38.	[ANONYMIZED]/danish-dynaword-prefix-continuation	Synthetic + audited	252M	0.35728%
39.	[ANONYMIZED]/danish-dynaword-span-filling	Synthetic + audited	251M	0.35593%
40.	[ANONYMIZED]/mtc1-synthetic-code-debugging	Synthetic + audited	247M	0.34996%
41.	GEM/wiki_cat_sum	Reformatted	205M	0.29117%
42.	allenai/tulu-3-sft-personas-math	Reformatted	204M	0.28922%
43.	giannor/gec_dala_tv2r_it	Reformatted	193M	0.27377%
44.	synquid/wildchat-100k-qwen-messages	Reformatted	190M	0.2694%
45.	[ANONYMIZED]/mtc1-synthetic-strict-math-answer-contract	Synthetic + audited	173M	0.24605%
46.	[ANONYMIZED]/mtc1-synthetic-constrained-format-following	Synthetic + audited	171M	0.24275%
47.	oliverkinch/danish-summarization	Reformatted	168M	0.23888%
48.	[ANONYMIZED]/common-pile-paragraph-reordering	Synthetic + audited	167M	0.23653%
49.	[ANONYMIZED]/danish-dynaword-paragraph-reordering	Synthetic + audited	160M	0.22769%
50.	glaiveai/glaive-function-calling-v2	Tool-call formatted	156M	0.22066%
51.	oliverkinch/machine-translation-da-ar	Reformatted	140M	0.19802%
52.	allenai/SciRIFF-train-mix	Reformatted	132M	0.18771%
53.	common-pile/arxiv_papers_filtered	Derived task	130M	0.18386%
54.	oliverkinch/da-instruct-dynaword	Reformatted	103M	0.14592%
55.	synquid/translation-100k	Reformatted	98.2M	0.13938%
56.	kobprof/skolegpt-instruct	Reformatted	86.4M	0.12259%
57.	allenai/verifiable-reasoning-filtered-o4-mini	Reformatted	80.2M	0.11378%
58.	oliverkinch/tidsskrift-dk-bt	Reformatted	76.7M	0.10878%
59.	allenai/open_math_2_50k_r1-original	Reformatted	71.8M	0.1019%
60.	oliverkinch/machine-translation-da-uk	Reformatted	69.5M	0.09859%
61.	giannor/dala_tv2r_it	Reformatted	68.5M	0.09723%
62.	Salesforce/xlam-function-calling-60k	Tool-call formatted	67.0M	0.09502%

continued from previous page

#	Dataset	Form	Tokens/epoch	Share
63.	oliverkinch/danish-qa	Reformatted	57.5M	0.08153%
64.	oliverkinch/dst-table-prompts-bt	Reformatted	51.1M	0.07248%
65.	danish-foundation-models/ai_arena_udtraek	Reformatted	45.7M	0.06483%
66.	oliverkinch/multi-wiki-qa-high-quality-subset	Reformatted	41.9M	0.05947%
67.	oliverkinch/dynaword-bt	Reformatted	34.3M	0.04867%
68.	HuggingFaceH4/no_robots	Reformatted	31.4M	0.04459%
69.	oliverkinch/da-instruct-dynaword-contemporary-hq	Reformatted	25.6M	0.03626%
70.	Team-ACE/ToolACE	Tool-call formatted	25.2M	0.03579%
71.	oliverkinch/da-instruct-dynaword-hq	Reformatted	24.1M	0.03414%
72.	oliverkinch/danish-university-portals-bt	Reformatted	21.8M	0.03088%
73.	allenai/tulu-3-sft-personas-algebra	Reformatted	21.0M	0.02982%
74.	allenai/IF_sft_data_verified	Reformatted	19.9M	0.02829%
75.	oliverkinch/autodata-da-sft	Reformatted	18.9M	0.02677%
76.	oliverkinch/danmarks-statistik-bt	Reformatted	18.9M	0.02676%
77.	synquid/danish-verifiable-reasoning	Reformatted	18.1M	0.02573%
78.	oliverkinch/da-instruct-dynaword-contemporary	Reformatted	17.6M	0.02491%
79.	oliverkinch/eur-lex-bt	Reformatted	16.7M	0.02369%
80.	allenai/tulu-3-sft-personas-code	Reformatted	14.6M	0.02077%
81.	oliverkinch/instruct-bt	Reformatted	13.5M	0.0191%
82.	[ANONYMIZED]/sapient-synth-flan-dialog-fsopt-data-qrecc	Synthetic + audited	13.1M	0.0186%
83.	synquid/ifbench-train	Reformatted	12.7M	0.018%
84.	allenai/tulu-3-sft-personas-instruction-following	Reformatted	11.5M	0.01634%
85.	[ANONYMIZED]/sapient-synth-flan-flan-fsopt-data-aeslc-1.0.0	Synthetic + audited	10.2M	0.01441%
86.	synquid/mt-da-deepseek	Reformatted	8.9M	0.01261%
87.	allenai/RLVR-MATH	Reformatted	8.1M	0.01145%
88.	allenai/RLVR-GSM	Reformatted	6.8M	0.00968%
89.	allenai/Dolci-Instruct-SFT-Tool-Use-SA	Tool-call formatted	5.8M	0.00825%
90.	[ANONYMIZED]/sapient-synth-flan-flan-fsnoopt-data-aeslc-1.0.0	Synthetic + audited	4.8M	0.0068%
91.	ccdv/govreport-summarization	Reformatted	4.4M	0.00626%
92.	[ANONYMIZED]/sapient-synth-flan-flan-fsopt-data-opinion-abstracts-rotten-tomatoes	Synthetic + audited	4.3M	0.00612%
93.	[ANONYMIZED]/sapient-synth-flan-dialog-fsopt-data-qrecc-ii	Synthetic + audited	2.9M	0.00413%
94.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task589-amazonfood-summary-text-generation	Synthetic + audited	2.8M	0.00399%
95.	oliverkinch/eur-lex-sum-instruct	Reformatted	2.6M	0.00373%
96.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task590-amazonfood-summary-correction-classification	Synthetic + audited	2.5M	0.0036%
97.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task618-amazonreview-summary-text-generation	Synthetic + audited	2.2M	0.00309%

continued from previous page

#	Dataset	Form	Tokens/epoch	Share
98.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1309-amazonreview-summary-classification	Synthetic + audited	1.9M	0.00276%
99.	[ANONYMIZED]/sapient-synth-flan-flan-zsnoopt-data-aeslc-1.0.0	Synthetic + audited	1.9M	0.00271%
100.	[ANONYMIZED]/sapient-synth-flan-flan-fsnoopt-data-opinion-abstracts-rotten-tomatoes	Synthetic + audited	1.7M	0.00247%
101.	[ANONYMIZED]/sapient-synth-flan-flan-zsopt-data-aeslc-1.0.0	Synthetic + audited	1.7M	0.00245%
102.	[ANONYMIZED]/sapient-synth-flan-dialog-zsopt-data-qrecc	Synthetic + audited	1.7M	0.00245%
103.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1375-newscomm-translation	Synthetic + audited	1.2M	0.00166%
104.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1373-newscomm-translation	Synthetic + audited	1.1M	0.0015%
105.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1376-newscomm-translation	Synthetic + audited	974K	0.00138%
106.	[ANONYMIZED]/sapient-synth-platypus-reclor	Synthetic + audited	922K	0.00131%
107.	[ANONYMIZED]/sapient-synth-tasksource-reclor	Synthetic + audited	901K	0.00128%
108.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task902-deceptive-opinion-spam-classification	Synthetic + audited	878K	0.00125%
109.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task634-allegro-reviews-classification	Synthetic + audited	816K	0.00116%
110.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task635-allegro-reviews-answer-generation	Synthetic + audited	816K	0.00116%
111.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task590-amazonfood-summary-correction-classification	Synthetic + audited	811K	0.00115%
112.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1309-amazonreview-summary-classification	Synthetic + audited	795K	0.00113%
113.	[ANONYMIZED]/sapient-synth-flan-flan-fsopt-data-opinion-abstracts-idebate	Synthetic + audited	746K	0.00106%
114.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1370-newscomm-classification	Synthetic + audited	721K	0.00102%
115.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task589-amazonfood-summary-text-generation	Synthetic + audited	705K	0.001%
116.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1371-newscomm-translation	Synthetic + audited	703K	0.001%
117.	oliverkinch/doab-da-bt	Reformatted	694K	0.00098%
118.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1377-newscomm-translation	Synthetic + audited	679K	0.00096%
119.	danish-foundation-models/kaenguruen	Reformatted	638K	0.0009%

continued from previous page

#	Dataset	Form	Tokens/epoch	Share
120.	[ANONYMIZED]/sapient-synth-flan-flan-zsopt-data-opinion-abstracts-rotten-tomatoes	Synthetic + audited	622K	0.00088%
121.	[ANONYMIZED]/sapient-synth-flan-flan-zsnoopt-data-opinion-abstracts-rotten-tomatoes	Synthetic + audited	616K	0.00087%
122.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task870-msmarco-answer-generation	Synthetic + audited	599K	0.00085%
123.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task1374-newscomm-translation	Synthetic + audited	599K	0.00085%
124.	[ANONYMIZED]/sapient-synth-tasksource-pragmeval-sarcasm	Synthetic + audited	542K	0.00077%
125.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task903-deceptive-opinion-spam-classification	Synthetic + audited	535K	0.00076%
126.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task618-amazonreview-summary-text-generation	Synthetic + audited	532K	0.00076%
127.	[ANONYMIZED]/sapient-synth-flan-flan-fsnoopt-data-opinion-abstracts-idebate	Synthetic + audited	527K	0.00075%
128.	[ANONYMIZED]/sapient-synth-flan-dialog-zsopt-data-qrecc-ii	Synthetic + audited	513K	0.00073%
129.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task265-paper-reviews-language-identification	Synthetic + audited	395K	0.00056%
130.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task635-allegro-reviews-answer-generation	Synthetic + audited	320K	0.00045%
131.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task909-dialogre-prevalent-speakers	Synthetic + audited	296K	0.00042%
132.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1376-newscomm-translation	Synthetic + audited	274K	0.00039%
133.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1373-newscomm-translation	Synthetic + audited	268K	0.00038%
134.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1375-newscomm-translation	Synthetic + audited	265K	0.00038%
135.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task266-paper-reviews-reviewer-perspective	Synthetic + audited	236K	0.00033%
136.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task906-dialogre-identify-names	Synthetic + audited	235K	0.00033%
137.	[ANONYMIZED]/sapient-synth-flan-flan-zsnoopt-data-opinion-abstracts-idebate	Synthetic + audited	223K	0.00032%
138.	[ANONYMIZED]/sapient-synth-flan-flan-zsopt-data-opinion-abstracts-idebate	Synthetic + audited	223K	0.00032%

continued from previous page

#	Dataset	Form	Tokens/epoch	Share
139.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task634-allegro-reviews-classification	Synthetic + audited	215K	0.0003%
140.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task870-msmarco-answer-generation	Synthetic + audited	212K	0.0003%
141.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task902-deceptive-opinion-spam-classification	Synthetic + audited	198K	0.00028%
142.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task672-amazon-yelp-summarization	Synthetic + audited	198K	0.00028%
143.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1377-newscomm-translation	Synthetic + audited	182K	0.00026%
144.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1374-newscomm-translation	Synthetic + audited	174K	0.00025%
145.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1371-newscomm-translation	Synthetic + audited	171K	0.00024%
146.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task265-paper-reviews-language-identification	Synthetic + audited	154K	0.00022%
147.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task903-deceptive-opinion-spam-classification	Synthetic + audited	149K	0.00021%
148.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task264-paper-reviews-accept-reject	Synthetic + audited	140K	0.0002%
149.	[ANONYMIZED]/sapient-synth-platypus-scibench	Synthetic + audited	104K	0.00015%
150.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task672-amazon-yelp-summarization	Synthetic + audited	85.3K	0.00012%
151.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task909-dialogre-prevalent-speakers	Synthetic + audited	82.0K	0.00012%
152.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task266-paper-reviews-reviewer-perspective	Synthetic + audited	71.9K	0.0001%
153.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task871-msmarco-question-generation	Synthetic + audited	71.1K	0.0001%
154.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task908-dialogre-identify-familial-relationships	Synthetic + audited	68.4K	0.0001%
155.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task906-dialogre-identify-names	Synthetic + audited	57.5K	0.00008%
156.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task264-paper-reviews-accept-reject	Synthetic + audited	49.3K	0.00007%
157.	[ANONYMIZED]/sapient-synth-flan-niv2-fsopt-data-task907-dialogre-identify-relationships	Synthetic + audited	46.0K	0.00007%

continued from previous page

#	Dataset	Form	Tokens/epoch	Share
158.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task871-msmarco-question-generation	Synthetic + audited	35.6K	0.00005%
159.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task908-dialogre-identify-familial-relationships	Synthetic + audited	17.3K	0.00002%
160.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task907-dialogre-identify-relationships	Synthetic + audited	13.4K	0.00002%
161.	[ANONYMIZED]/sapient-synth-flan-niv2-zsopt-data-task1370-newscomm-classification	Synthetic + audited	1.9K	<0.00001%

F Evaluation Setup and Datasets

All models used vLLM-served (Kwon et al., 2023) endpoints with FlashInfer, with the exception of Mimir, which requires FlashAttention to correctly capture the PrefixLM and Gemma 4 chat template. We ran both vLLM with FlashAttention4 (Zadouri et al., 2026) and Hugging Face Transformers, obtaining comparable results up to numerical stability.

For ease of reproduction, we report the results from Hugging Face Transformers (Wolf et al., 2019).

Gemma 4 was evaluated in both non-thinking and thinking modes. Thinking requires ∼500–650 tokens. MCQ tasks use `max_tokens=1`. All non-MCQ tasks (or whenever reasoning is enabled) use `max_tokens=2048`. All baseline evaluations were conducted via the Inspect AI Framework of the AI Security Institute (2024).

Table 14 lists all considered evaluation datasets with their precise source as Hugging Face identifier, and the number of in-context shots provided to the models during evaluation. We use the few-shot configuration from Wang et al. (2026) the English evaluation datasets, while we use zero-shot evaluation for Math & Code as well as Danish evaluation datasets.

Table 14: Overview of benchmarks and their configuration. All benchmarks are evaluated with *temperature* 0 and a fixed seed.

Task	Dataset (split)	N-shots	Reference
English			
BoolQ	google/boolq	5	Clark et al. (2019)
Winogrande	allenai/winogrande	5	Sakaguchi et al. (2021)
Hellaswag	Rowan/hellaswag	10	Zellers et al. (2019)
MMLU	cais/mmlu	5	Hendrycks et al. (2021a)
ARC-C	allenai/ai2_arc	25	Clark et al. (2018)
DROP	EleutherAI/drop	3	Dua et al. (2019)
GovReport	ccdv/govreport	0	Huang et al. (2021)
Math & Code			
GSM8K	openai/gsm8k	0	Cobbe et al. (2021)
MATH	EleutherAI/hendrycks_math	0	Hendrycks et al. (2021b)
HumanEval	human_eval (164 problems)	0	Chen et al. (2021)
Danish			
Angry Tweets	DDSC/angry-tweets	0	Pauli et al. (2021)
DaLA	giannor/dala	0	Barmina et al. (2026a)
GEC-DaLA	giannor/dala_gen_v3	0	Barmina et al. (2026a)
PIQA-da	local JSON	0	Chang et al. (2026)
Daisy	[ANONYMIZED]/SDU-Daisy	0	Nielsen et al. (2026)
Multi Wiki QA	oliverkinch/multi-wiki-qa	0	Smart (2026)
WMT24++ EN-DA	synquid/wmt24pp	0	Deutsch et al. (2025)
Nordj. News Summ.	alexandrainst/nordjylland-news-summarization	0	Kinch (2023)
IFEval-Da	danish-foundation-models/ifeval-da	0	Larsen et al. (2025)
Hellaswag-da	EuroEval	0	Nielsen (2023)