

# Zen-Pro: A High-Performance Language Model for Complex Reasoning

Technical Report v2025.02

Zen LM Research Team  
research@zenlm.org

February 2025

## Abstract

We present **Zen-Pro**, a 72 billion parameter dense transformer model designed for complex reasoning, professional scientific applications, and advanced code generation. Zen-Pro extends the Zen MoDE (Mixture of Distilled Experts) architecture to the 72B parameter scale, trained on 4.5 trillion tokens with a curriculum emphasizing mathematics, science, and multi-step reasoning. Post-training combines extended chain-of-thought (CoT) supervised fine-tuning with reinforcement learning from verifiable rewards (RLVR) on formally checkable problem domains. Zen-Pro achieves frontier-level performance: MMLU 89.2%, MATH 84.1%, HumanEval 91.3%, and GPQA 71.8%, competitive with the strongest models at this capability tier while supporting a 128K token context window for long-form professional workflows.

## Contents

|          |                                                |          |
|----------|------------------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                            | <b>3</b> |
| <b>2</b> | <b>Architecture</b>                            | <b>3</b> |
| 2.1      | Model Hyperparameters                          | 3        |
| 2.2      | Grouped Query Attention at Scale               | 3        |
| 2.3      | Sliding Window Attention for Long Context      | 4        |
| 2.4      | Extended Context via RoPE Scaling              | 4        |
| <b>3</b> | <b>Training Methodology</b>                    | <b>4</b> |
| 3.1      | Pretraining Data Curriculum                    | 4        |
| 3.2      | Long Chain-of-Thought Fine-Tuning              | 5        |
| 3.3      | Reinforcement Learning from Verifiable Rewards | 5        |
| <b>4</b> | <b>Evaluation</b>                              | <b>5</b> |
| 4.1      | Standard Benchmarks                            | 5        |
| 4.2      | Scientific Reasoning                           | 6        |
| 4.3      | Long-Context Evaluation                        | 6        |
| 4.4      | Coding Benchmarks                              | 6        |
| 4.5      | Inference Efficiency                           | 6        |
| <b>5</b> | <b>Ablation Studies</b>                        | <b>7</b> |
| 5.1      | Effect of RLVR                                 | 7        |
| 5.2      | Effect of Long CoT Training                    | 7        |
| <b>6</b> | <b>Related Work</b>                            | <b>7</b> |

|   |             |   |
|---|-------------|---|
| 7 | Limitations | 7 |
| 8 | Conclusion  | 8 |

# 1 Introduction

The transition from 7B to 70B+ parameter models represents more than a quantitative scaling step. Empirical research consistently shows qualitative capability emergence at larger scales, including multi-step mathematical reasoning, scientific hypothesis evaluation, and nuanced instruction disambiguation [1, 2]. Zen-Pro is our 72B offering in this tier, designed to serve professional and enterprise use cases where accuracy, reasoning depth, and reliability are paramount.

Zen-Pro makes the following contributions:

- A 72B dense transformer trained on 4.5T tokens with a curriculum that front-loads high-quality mathematical, scientific, and code data in the final training phase.
- Extended post-training with long chain-of-thought (CoT) fine-tuning and reinforcement learning from verifiable rewards (RLVR) using formal verification on math and code tasks.
- A 128K token context window via extended RoPE, validated on real-world long-document tasks including legal review, codebase analysis, and scientific literature synthesis.
- Frontier-tier benchmark performance that closes the gap with models of significantly larger total parameter counts.

Zen-Pro serves as the flagship high-performance model in the Zen family for deployments requiring production-grade reasoning without the infrastructure overhead of the Zen-Max MoE tier.

## 2 Architecture

### 2.1 Model Hyperparameters

Zen-Pro scales the Zen MoDE architecture to 72B parameters. Table 1 lists the key architectural hyperparameters.

Table 1: Zen-Pro architecture hyperparameters.

| Hyperparameter             | Value                         |
|----------------------------|-------------------------------|
| Parameters (total)         | 72.7B                         |
| Layers                     | 80                            |
| Attention heads            | 64                            |
| KV heads (GQA)             | 8                             |
| Hidden dimension           | 8192                          |
| FFN intermediate dimension | 29,568                        |
| Vocabulary size            | 151,936                       |
| Context length (training)  | 131,072                       |
| Position encoding          | RoPE ( $\theta = 5,000,000$ ) |
| Activation function        | SiLU                          |
| Normalization              | RMSNorm                       |
| Tied embeddings            | No                            |

### 2.2 Grouped Query Attention at Scale

At 72B parameters and 80 layers, memory bandwidth is the dominant inference bottleneck. Zen-Pro uses GQA with 64 query heads and 8 KV heads, achieving an 8 $\times$  reduction in KV cache size relative to full multi-head attention (MHA). This enables:

$$\text{KV cache size} = 2 \cdot L \cdot n_{\text{kv}} \cdot d_k \cdot S \cdot \text{dtype\_bytes} \quad (1)$$

where  $L = 80$  layers,  $n_{\text{kv}} = 8$  KV heads,  $d_k = 128$  head dimension, and  $S$  is sequence length. For a 128K-token sequence in BF16, this yields approximately 26.8 GB of KV cache, feasible on  $4 \times$  A100-80GB with tensor parallelism.

### 2.3 Sliding Window Attention for Long Context

For the final 20 layers (layers 60–79), Zen-Pro employs sliding window attention (SWA) [3] with a window size of 16K tokens in addition to full global attention on every 4th layer. This hybrid pattern maintains linear scaling in memory for long sequences while preserving global information flow:

$$\text{Attention scope} = \begin{cases} \text{full } [0, S] & \text{if } l \bmod 4 = 0 \\ \text{window } [i - w, i] & \text{otherwise} \end{cases} \quad (2)$$

where  $w = 16,384$  is the sliding window size and  $i$  is the current token position.

### 2.4 Extended Context via RoPE Scaling

To support 128K tokens, we use a higher RoPE base frequency of  $\theta = 5,000,000$  combined with position interpolation during the long-context fine-tuning phase. Pretraining is conducted at 8K context, with a two-stage context extension:

1. Extend to 32K at reduced learning rate ( $5 \times 10^{-5}$ ) for 10B tokens.
2. Extend to 128K at further reduced learning rate ( $1 \times 10^{-5}$ ) for 5B tokens.

This staged extension preserves short-context performance while enabling long-context generalization.

## 3 Training Methodology

### 3.1 Pretraining Data Curriculum

Zen-Pro trains on 4.5T tokens with a data curriculum that emphasizes reasoning-intensive sources. The final 10% of training uses a concentrated curriculum of the highest-quality sources.

Table 2: Pretraining data composition (4.5T tokens total).

| Domain                        | Tokens (B) | Fraction |
|-------------------------------|------------|----------|
| Web text (quality-filtered)   | 1,800      | 40.0%    |
| Books and long-form           | 675        | 15.0%    |
| Code (all languages)          | 675        | 15.0%    |
| Scientific articles           | 450        | 10.0%    |
| Mathematics (formal+informal) | 360        | 8.0%     |
| Multilingual web              | 315        | 7.0%     |
| Curated reasoning chains      | 225        | 5.0%     |
| Total                         | 4,500      | 100.0%   |

Curated reasoning chains include synthetic chain-of-thought solutions to mathematical competition problems (AMC, AIME, Olympiad), formal theorem proving datasets (Lean, Isabelle), and verified code problem solutions with test suite passes.

### 3.2 Long Chain-of-Thought Fine-Tuning

After base pretraining, we apply a specialized CoT-SFT phase on 2.5M long-form reasoning trajectories. Unlike standard SFT which targets short responses, CoT-SFT trains the model to generate step-by-step reasoning of 500–5,000 tokens before producing a final answer:

$$\mathcal{L}_{\text{CoT-SFT}} = - \sum_{t=1}^{T_r} \log p(r_t \mid x, r_{<t}) - \lambda \sum_{t=1}^{T_a} \log p(a_t \mid x, r, a_{<t}) \quad (3)$$

where  $r$  is the reasoning chain,  $a$  is the final answer, and  $\lambda = 0.5$  down-weights the answer loss to encourage richer intermediate reasoning.

### 3.3 Reinforcement Learning from Verifiable Rewards

RLVR [4, 5] replaces human preference labels with programmatically verifiable correctness signals on three domains:

- **Mathematics:** symbolic equivalence checking of numeric and algebraic answers.
- **Code:** unit test execution with pass@k evaluation.
- **Logic:** SAT/SMT solver verification of formal claims.

The reward function is:

$$r(y, y^*) = \mathbb{I}[\text{verify}(y) = y^*] - \alpha \cdot \frac{|y|}{L_{\max}} \quad (4)$$

where the first term is 1 for correct verified answers and the second term penalizes unnecessary verbosity with coefficient  $\alpha = 0.02$  and maximum allowed length  $L_{\max}$ .

Group Relative Policy Optimization (GRPO) [6] is used as the RL algorithm, sampling  $G = 8$  responses per prompt and computing advantages within the group to reduce variance without a separate value network.

## 4 Evaluation

### 4.1 Standard Benchmarks

Table 3: Zen-Pro benchmark results versus comparable frontier models.

| Benchmark           | Zen-Pro (72B) | Competitor A (70B) | Competitor B (72B) | Competitor C (70B) |
|---------------------|---------------|--------------------|--------------------|--------------------|
| MMLU (5-shot)       | <b>89.2</b>   | 88.1               | 87.6               | 82.5               |
| MMLU-Pro            | <b>72.4</b>   | 71.1               | 70.8               | 61.5               |
| ARC-Challenge       | 72.8          | 73.1               | 71.4               | 66.5               |
| HellaSwag           | 88.4          | 87.9               | 88.1               | 86.5               |
| WinoGrande          | 82.3          | 81.8               | 81.4               | 78.5               |
| MATH (4-shot, CoT)  | <b>84.1</b>   | 80.3               | 78.9               | 62.5               |
| GSM8K (8-shot, CoT) | <b>93.4</b>   | 92.1               | 91.8               | 87.5               |
| GPQA Diamond        | <b>71.8</b>   | 68.4               | 66.9               | 51.5               |
| HumanEval (pass@1)  | <b>91.3</b>   | 88.7               | 87.4               | 74.5               |
| MBPP (pass@1)       | <b>84.6</b>   | 82.3               | 80.9               | 71.5               |
| MT-Bench (1-10)     | <b>9.1</b>    | 8.9                | 8.7                | 8.5                |

## 4.2 Scientific Reasoning

Zen-Pro shows strong performance on scientific reasoning benchmarks, reflecting the emphasis on scientific literature and formal reasoning in the training curriculum.

Table 4: Scientific reasoning benchmark results.

| Benchmark             | Zen-Pro (72B) | Prior Best (Open) |
|-----------------------|---------------|-------------------|
| GPQA (Graduate-level) | 71.8          | 68.4              |
| GPQA (Diamond subset) | 64.3          | 61.7              |
| SciQ                  | 97.1          | 96.8              |
| ARC-Challenge         | 72.8          | 73.1              |
| MMLU-STEM subjects    | 91.4          | 89.2              |
| MedQA (USMLE)         | 82.6          | 80.1              |

## 4.3 Long-Context Evaluation

Table 5: RULER long-context benchmark scores at multiple context lengths.

| Model              | 8K   | 16K  | 32K  | 64K  | 128K |
|--------------------|------|------|------|------|------|
| Zen-Pro (72B)      | 96.2 | 94.8 | 92.1 | 88.3 | 82.7 |
| Competitor A (70B) | 95.8 | 92.3 | 86.4 | 71.2 | 52.3 |
| Competitor B (72B) | 94.9 | 91.7 | 85.1 | 68.4 | 48.7 |

## 4.4 Coding Benchmarks

Table 6: Code generation benchmark results.

| Benchmark                   | Zen-Pro (72B) | Competitor A (70B) |
|-----------------------------|---------------|--------------------|
| HumanEval (pass@1)          | 91.3          | 88.7               |
| HumanEval+ (pass@1)         | 87.1          | 84.2               |
| MBPP (pass@1)               | 84.6          | 82.3               |
| MBPP+ (pass@1)              | 79.3          | 76.8               |
| SWE-bench Verified (pass@1) | 38.7          | 34.1               |
| LiveCodeBench (3-month)     | 62.4          | 58.3               |

## 4.5 Inference Efficiency

Table 7: Inference throughput on 4× A100-80GB (tensor parallel, BF16).

| Metric                      | 8K context | 128K context |
|-----------------------------|------------|--------------|
| Throughput (tok/s, batch=1) | 42         | 18           |
| Throughput (tok/s, batch=8) | 310        | 124          |
| TTFT P50 (ms, 4K prompt)    | 210        | 1,840        |
| GPU memory (weights+KV)     | 148 GB     | 174 GB       |

## 5 Ablation Studies

### 5.1 Effect of RLVR

Table 8 compares Zen-Pro with RLVR against the SFT-only checkpoint.

Table 8: Ablation: effect of RLVR post-training stage.

| Model                    | MATH | HumanEval | GPQA |
|--------------------------|------|-----------|------|
| Zen-Pro SFT only         | 77.3 | 86.2      | 63.4 |
| Zen-Pro + RLVR           | 84.1 | 91.3      | 71.8 |
| Improvement ( $\Delta$ ) | +6.8 | +5.1      | +8.4 |

RLVR produces the largest gains on GPQA, reflecting that formal scientific reasoning benefits most from verifiable reward signals where human preference annotation is difficult to obtain at scale and quality.

### 5.2 Effect of Long CoT Training

Table 9: Ablation: standard SFT vs. long chain-of-thought SFT.

| Model           | MATH | AIME 2024 |
|-----------------|------|-----------|
| Standard SFT    | 71.4 | 23.3%     |
| Long CoT SFT    | 78.9 | 36.7%     |
| Long CoT + RLVR | 84.1 | 53.3%     |

## 6 Related Work

Large-scale dense transformers at the 70B tier have been explored extensively following the LLaMA 2 [9] release, with subsequent models pushing reasoning capability through data quality improvements and specialized post-training. Reinforcement learning with verifiable rewards traces to outcome-based reward modeling [5] and was prominently applied in mathematical reasoning work [4, 6]. Our GRPO implementation follows the group relative policy gradient approach, adapted for a mix of math, code, and logic verification environments.

Long-context scaling via RoPE interpolation has been studied in multiple works [7, 8], and the hybrid sliding-window / full-attention pattern draws inspiration from Longformer [3]. Our two-stage context extension is a pragmatic variant of these approaches optimized for training stability.

## 7 Limitations

Zen-Pro at 72B requires significant compute for inference: at minimum  $2\times$  A100-80GB in BF16 without quantization, limiting accessibility compared to smaller models. While RLVR substantially improves mathematical and code reasoning, commonsense reasoning and open-ended creative tasks rely entirely on SFT signal, leaving room for improvement. Long-context performance degrades for retrieval tasks beyond 64K tokens, and generation quality on very long outputs ( $>8K$  tokens) has not been systematically evaluated.

## 8 Conclusion

Zen-Pro represents the Zen family’s high-performance tier, demonstrating that rigorous data curation combined with verifiable-reward reinforcement learning yields frontier-level reasoning capability at the 72B scale. The model achieves MMLU 89.2%, MATH 84.1%, HumanEval 91.3%, and GPQA 71.8%, competitive with the strongest open-weight models while maintaining a 128K context window for professional workflows. The RLVR training protocol generalizes well across verifiable domains and is expected to underpin future Zen family post-training pipelines.

## References

- [1] J. Wei et al., “Emergent Abilities of Large Language Models,” *TMLR*, 2022.
- [2] A. Srivastava et al., “Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models,” *arXiv:2206.04615*, 2022.
- [3] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The Long-Document Transformer,” *arXiv:2004.05150*, 2020.
- [4] H. Lightman et al., “Let’s Verify Step by Step,” *arXiv:2305.20050*, 2023.
- [5] K. Cobbe et al., “Training Verifiers to Solve Math Word Problems,” *arXiv:2110.14168*, 2021.
- [6] Z. Shao et al., “DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models,” *arXiv:2402.03300*, 2024.
- [7] B. Peng et al., “YaRN: Efficient Context Window Extension of Large Language Models,” *arXiv:2309.00071*, 2023.
- [8] S. Chen et al., “Extending Context Window of Large Language Models via Positional Interpolation,” *arXiv:2306.15595*, 2023.
- [9] H. Touvron et al., “Llama 2: Open Foundation and Fine-Tuned Chat Models,” *arXiv:2307.09288*, 2023.
- [10] J. Su et al., “RoFormer: Enhanced Transformer with Rotary Position Embedding,” *arXiv:2104.09864*, 2021.
- [11] J. Ainslie et al., “GQA: Training Generalized Multi-Query Transformer Models,” *EMNLP*, 2023.
- [12] R. Rafailov et al., “Direct Preference Optimization,” *NeurIPS*, 2023.
- [13] L. Ouyang et al., “Training Language Models to Follow Instructions with Human Feedback,” *NeurIPS*, 2022.