

# Zen: Ultra-Efficient Language Models for Local Deployment and Privacy Preservation

\*Technical Report v1.0.1

Hanzo AI Research Team<sup>1</sup>

Zoo Labs Foundation<sup>2</sup>

<sup>1</sup>Hanzo AI (Techstars '17)

Email: research@hanzo.ai

<sup>2</sup>Zoo Labs Foundation (501(c)(3))

Email: foundation@zoo.ai

**Abstract**—The Zen model family represents a breakthrough in efficient AI deployment, achieving state-of-the-art performance while reducing computational requirements by up to 98%. This technical report provides an overview of the Zen architecture, training methodology, and deployment strategies. We demonstrate that with careful architecture design and optimization, models ranging from 0.6B to 480B parameters can be deployed across diverse hardware platforms from edge devices to cloud infrastructure, democratizing access to frontier AI capabilities while maintaining strong performance on standard benchmarks.

## I. INTRODUCTION

The exponential growth in AI model capabilities has been accompanied by equally dramatic increases in computational requirements. The Zen model family addresses this challenge through a principled approach to model design that prioritizes efficiency without compromising capability.

Our key contributions include:

- A family of 10 models spanning language, vision, and audio modalities
- Mixture-of-Experts architectures that activate only 10-20% of parameters
- Extended thinking modes supporting up to 2M internal reasoning tokens
- Deployment formats supporting 4-bit quantization with minimal quality loss
- Environmental impact reduction of up to 98% compared to equivalent models

## II. MODEL ARCHITECTURE

The Zen family comprises models built on modern transformer architectures with several key innovations:

### A. Language Models

**Zen-Nano (0.6B):** Optimized for edge deployment with grouped-query attention and INT4 quantization, achieving 51.7% MMLU while running at 450 tokens/sec on mobile devices.

**Zen-Eco (4B):** Balanced for consumer hardware with Flash Attention v2, supporting 32K context with 128K thinking tokens.

**Zen-Omni (30B):** Unified multimodal transformer with cross-modal attention for native text-image understanding.

**Zen-Coder (480B MoE, 30B active):** Specialized for code with 16 experts, 2 active per token, achieving 72.8% HumanEval.

**Zen-Next (80B):** Flagship dense model with 128K context and 1M thinking tokens for maximum capability.

### B. Visual Models

**Zen-Artist (8B):** Diffusion-based text-to-image generation up to  $1024 \times 1024$  resolution.

**Zen-Designer (235B MoE, 22B active):** Vision-language models for design analysis and generation with 2M thinking tokens.

### C. Audio Models

**Zen-Scribe (1.5B):** CTC/attention hybrid for 98-language speech recognition with 3.2% WER.

## III. TRAINING METHODOLOGY

Models are trained on a carefully curated corpus of 7T tokens with domain-specific augmentation. The training pipeline includes:

- 1) Pretraining on filtered web-scale data
- 2) Supervised fine-tuning on instruction datasets
- 3) RLHF with 10M preference comparisons
- 4) Constitutional AI for safety alignment
- 5) Quantization-aware fine-tuning for deployment

## IV. RESULTS

Model	MMLU	HumanEval	GSM8K
Zen-Nano	51.7	22.6	62.0
Zen-Eco	62.3	35.2	74.8
Zen-Omni	68.4	48.3	82.1
Zen-Coder	78.9	72.8	94.7
Zen-Next	75.6	61.7	90.7

TABLE I  
LANGUAGE MODEL BENCHMARK RESULTS (%)

## V. CONCLUSION

The Zen model family demonstrates that efficiency and capability are not mutually exclusive. Through careful architecture design, training optimization, and quantization techniques, we achieve state-of-the-art performance while reducing computational requirements by up to 98%, enabling deployment across diverse hardware platforms.

## ACKNOWLEDGMENTS

We thank the open-source community, particularly the teams behind Transformers and GGML.

## REFERENCES

- [1] Fedus et al., “Switch Transformers,” JMLR, 2022.