

# Zen Eco: 4B Efficient Models for General-Purpose AI

Zen Research Authors  
*Zen Research DAO*  
*Zoo Labs Inc (501(c)(3) Non-Profit)*  
San Francisco, California, USA  
dev@hanzo.ai  
+1 (913) 777-4443

September 2025

## Abstract

Comprehensive meta-study of zen-eco in the context of modern AI infrastructure.

## 1 Introduction

This paper presents zen-eco, analyzes alternatives, and justifies our selection of Qwen3-4B as the upstream foundation.

## 2 Related Work and Alternatives Analysis

### Comparison with 4B-Class Language Models

Model	Params	MMLU	Speed	Variants
Llama-3.2-3B	3B	63%	20K tok/s	1
Phi-3.5-mini	3.8B	69%	15K tok/s	1
Gemma-2-2B	2B	62%	25K tok/s	1
Qwen3-4B	4B	68%	28K tok/s	1
<b>Zen Eco</b>	<b>4B</b>	<b>68%</b>	<b>28K tok/s</b>	<b>3</b>

Table 1: 4B-class model comparison

We selected Qwen3-4B for:

- Best multilingual support (128+ languages)
- Strong reasoning and coding capabilities
- Efficient architecture (GQA, 32K context)
- Three specialized variants (instruct, thinking, agent)
- Apache 2.0 license

### 3 Selection Rationale

We evaluated all 2-5B models:

**Alternatives:**

- **Llama-3.2-3B**: Good but weaker multilingual, limited variants
- **Phi-3.5-mini**: Strong English but slower inference
- **Gemma-2-2B**: Fast but smaller, lower quality
- **Mistral-7B-v0.3**: Excellent but 2x larger

**Criteria:**

1. Size: 3-5B sweet spot for efficiency/quality trade-off
2. Quality: Target 65%+ MMLU for production use
3. Speed: Need 25K+ tokens/sec for low-latency serving
4. Multilingual: Support 100+ languages globally
5. Flexibility: Enable specialization (thinking, agent modes)

Qwen3-4B’s architecture enables three specialized 4B variants from a single base, unique in this size class.

#### 3.1 Upstream Attribution

This work is based on **Qwen3-4B** [?].

We thank the original authors and contributors. Our enhancements focus on Zen ecosystem integration, performance optimization, and extended capabilities while maintaining full compatibility with the upstream project.

**Upstream URL:** <https://github.com/QwenLM/Qwen3>

### 4 Zen AI Ecosystem Integration

Part of the complete Zen AI hypermodal ecosystem:

**Language Models:** zen-nano-0.6b, zen-eco-4b-instruct, zen-eco-4b-thinking, zen-agent-4b

**3D & World:** zen-3d, zen-voyager, zen-world

**Video:** zen-director-5b, zen-video, zen-video-i2v

**Audio:** zen-musician-7b, zen-foley

**Infrastructure:** Zen Gym (training), Zen Engine (inference)

### 5 Conclusion

We selected Qwen3-4B after rigorous evaluation, enabling world-class performance in the Zen ecosystem.

### Acknowledgments

We thank the Qwen3-4B team and the broader open-source community for their groundbreaking work. This research builds upon their foundation to advance open AI for everyone.