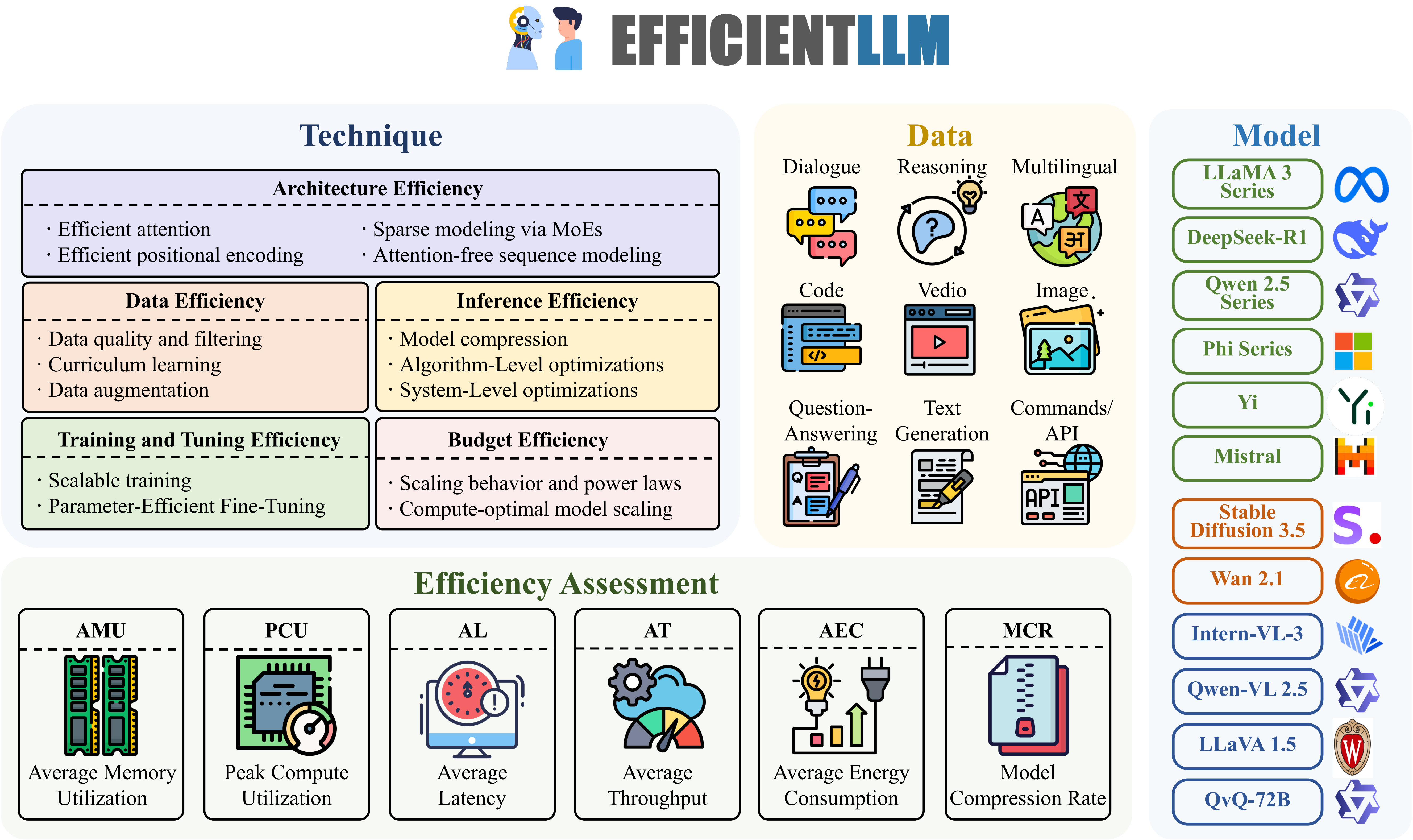

EFFICIENTLLM

Technique

Architecture Efficiency

- Efficient attention
- Efficient positional encoding
- Sparse modeling via MoEs
- Attention-free sequence modeling

Data Efficiency

- Data quality and filtering
- Curriculum learning
- Data augmentation

Inference Efficiency

- Model compression
- Algorithm-Level optimizations
- System-Level optimizations

Training and Tuning Efficiency

- Scalable training
- Parameter-Efficient Fine-Tuning

Budget Efficiency

- Scaling behavior and power laws
- Compute-optimal model scaling

Data

Model

Efficiency Assessment