- 2 Preliminaries
  - 2.1 Basic Architecture
    - 2.1.1 Typical Examples
    - 2.1.2 Architectures
      - Attention
      - RoPE
      - RMSNorm
      - SwiGLU/GeGLU
    - 2.1.3 Model Capability
    - 2.1.4 Deployment Efficiency
  - 2.2 Metrics
    - 2.2.1 Speed
      - TTFT
      - TBT
    - 2.2.2 Storage
      - RAM/VRAM
      - Flash Memory
    - 2.2.3 Energy
      - Battery Drain
      - Heating Problem
    - 2.2.4 Capability
      - Factuality
      - Long-Context
      - Math and Science
      - Reasoning
      - Multilingual
      - Medicine
      - Autonomous Driving
      - Current Benchmark
      - Limitations: Vertical Application / Multimodal Collaboration
  - 2.3 Edge Devices
    - 2.3.1 CPU
    - 2.3.2 Mobile GPU
    - 2.3.3 NPU
    - 2.3.4 PIM & PNM
    - 2.3.5 ASIC & FPGA