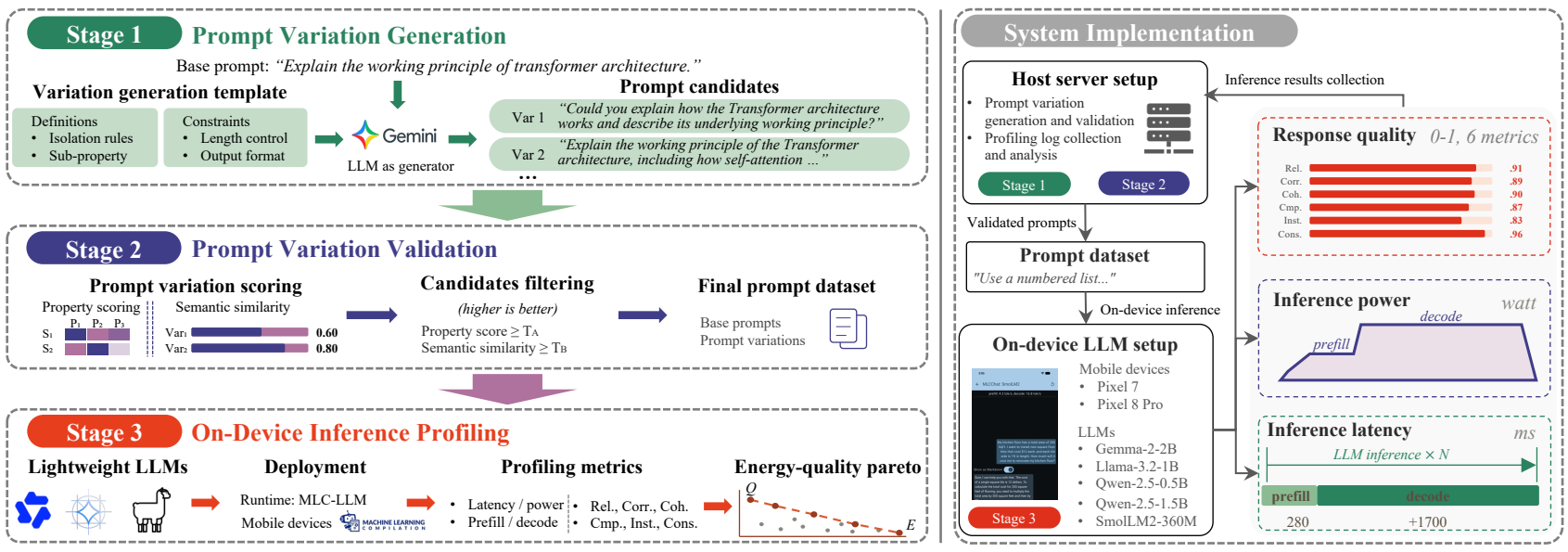
Stage 1
Prompt Variation Generation
Base prompt: "Explain the working principle of transformer architecture."
Variation generation template
Definitions
• Isolation rules
• Sub-property
Constraints
• Length control
• Output format
Gemini
LLM as generator
Prompt candidates
Var 1
"Could you explain how the Transformer architecture works and describe its underlying working principle?"
Var 2
"Explain the working principle of the Transformer architecture, including how self-attention ..."
...
Stage 2
Prompt Variation Validation
Prompt variation scoring
Property scoring
Semantic similarity
P1 P2 P3
S1
S2
Var1 0.60
Var2 0.80
Candidates filtering
(higher is better)
Property score ≥ TA
Semantic similarity ≥ TB
Final prompt dataset
Base prompts
Prompt variations
Stage 3
On-Device Inference Profiling
Lightweight LLMs
Deployment
Runtime: MLC-LLM
Mobile devices
MACHINE LEARNING COMPILATION
Profiling metrics
• Latency / power
• Prefill / decode
• Rel., Corr., Coh.
• Cmp., Inst., Cons.
Energy-quality pareto
Q
E
System Implementation
Host server setup
• Prompt variation generation and validation
• Profiling log collection and analysis
Stage 1
Stage 2
Inference results collection
Validated prompts
Prompt dataset
"Use a numbered list..."
On-device inference
On-device LLM setup
Mobile devices
• Pixel 7
• Pixel 8 Pro
LLMs
• Gemma-2-2B
• Llama-3.2-1B
• Qwen-2.5-0.5B
• Qwen-2.5-1.5B
• SmolLM2-360M
Stage 3
Response quality
0-1, 6 metrics
Rel. .91
Corr. .89
Coh. .90
Cmp. .87
Inst. .83
Cons. .96
Inference power
watt
decode
prefill
Inference latency
ms
LLM inference × N
prefill
decode
280
+1700