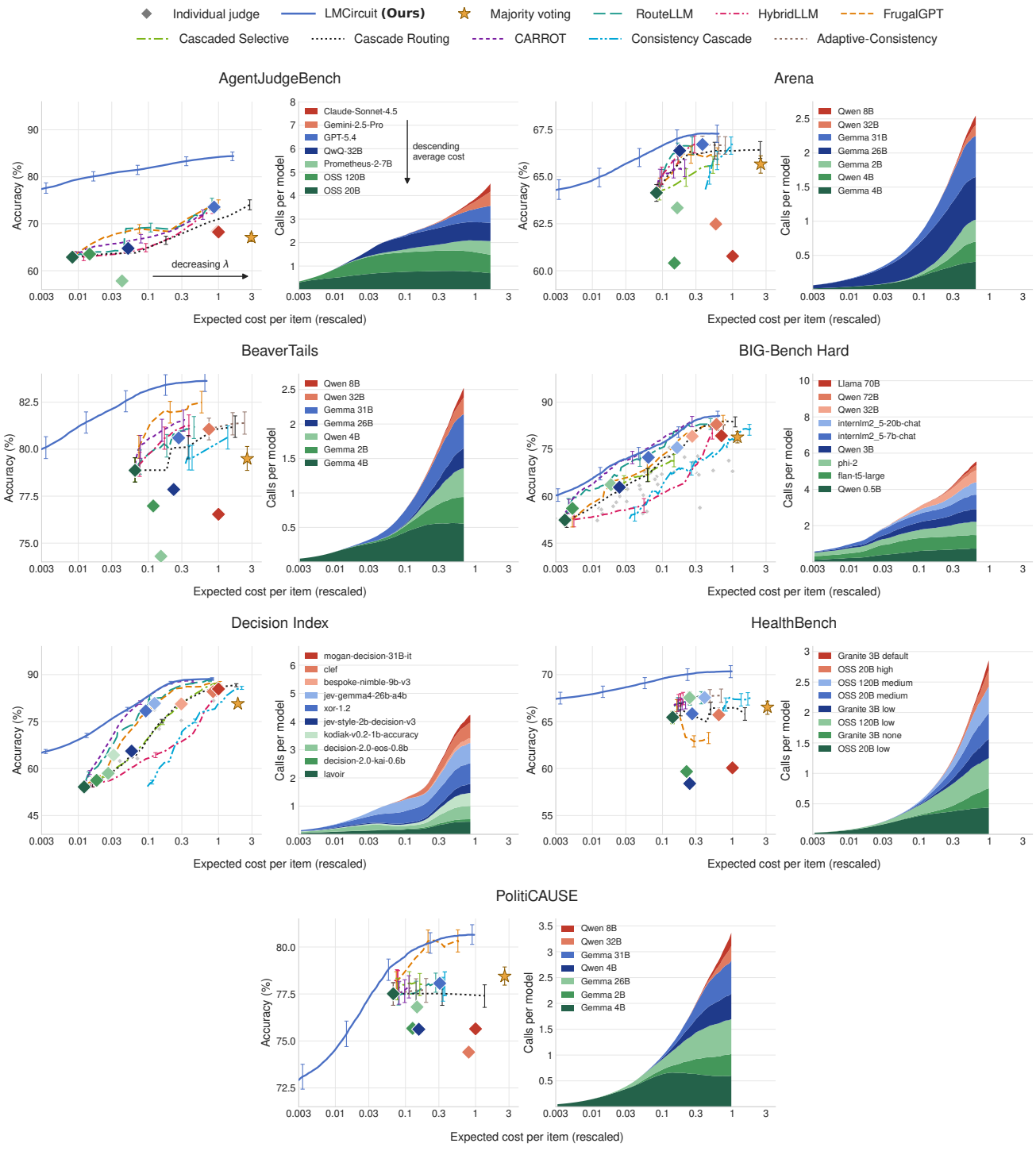

Individual judge
LMCircuit (Ours)
Majority voting
RouteLLM
HybridLLM
FrugalGPT
Cascaded Selective
Cascade Routing
CARROT
Consistency Cascade
Adaptive-Consistency
AgentJudgeBench
Arena
BeaverTails
BIG-Bench Hard
Decision Index
HealthBench
PolitiCAUSE
Accuracy (%)
Calls per model
Expected cost per item (rescaled)
decreasing λ
descending average cost
Claude-Sonnet-4.5
Gemini-2.5-Pro
GPT-5.4
QwQ-32B
Prometheus-2-7B
OSS 120B
OSS 20B
Qwen 8B
Qwen 32B
Gemma 31B
Gemma 26B
Gemma 2B
Qwen 4B
Gemma 4B
Llama 70B
Qwen 72B
internlm2_5-20b-chat
internlm2_5-7b-chat
Qwen 3B
phi-2
flan-t5-large
Qwen 0.5B
mogan-decision-31B-it
clef
bespoke-nimble-9b-v3
jev-gemma4-26b-a4b
xor-1.2
jev-style-2b-decision-v3
kodiak-v0.2-1b-accuracy
decision-2.0-eos-0.8b
decision-2.0-kai-0.6b
lavoir
Granite 3B default
OSS 20B high
OSS 120B medium
OSS 20B medium
Granite 3B low
OSS 120B low
Granite 3B none
OSS 20B low