面壁智能（ModelBest）：エッジAIの革新とMiniCPMの市場競争力

パラダイムシフト：クラウドからエッジへ

クラウド依存：蠏大な計算資源、運延発生、データ外部送信

デバイス完結：NPU/GPU推論、リアルタイム応答、安全性担保

技術的ブレイクスルーと効率性

モデル風洞実験
小規模モデルで最速学習率特定、大規模訓練コスト最小化

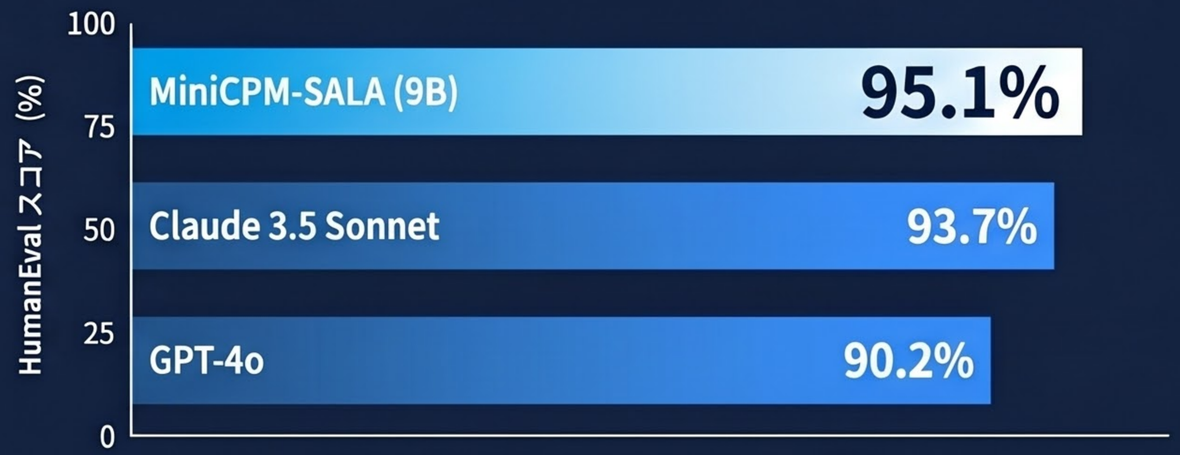
動画トークンを96倍に圧縮
(3D-Resampler)：動画を空間・時間的解析、情報冥総持でリアルタイム動画理解

1Mトークンの長文脈処理
MiniCPM-SALA：ハイブリッド注意機構で100万トークン推論、従来比3.5倍速度

Llama-3.1/Phi-3.5の約8倍のコンテキスト長、ローカル運用でAPIコストゼロ

指標	MiniCPM-SALA（9B）	Llama-3.1 (8B)	Phi-3.5-mini (3.8B)
最大コンテキスト長	1,000,000 (1M)	131,072 (128K)	128,000 (128K)
HumanEval（コード）	96.1%	72.6%	62.8%
APIコスト（1Mトークン）	ローカル主体（0円）	$0.03	$0.10

定量的パフォーマンス比較：HumanEval（コード生成）

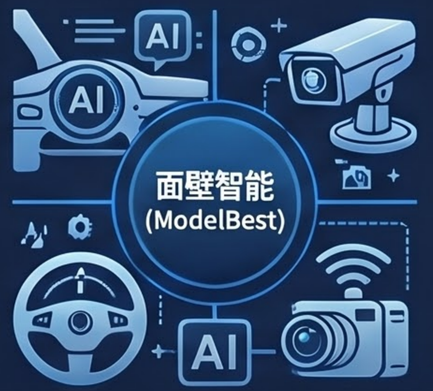


競合モデルを凌駕する推論能力を実証

戦略的パートナーシップと社会実装

国家インフラと実業融合
中国電信との提携、可決・業育分野へのAI導入提進

スマートデバイス
Huawei、Lenovo、PC・スマホ実装



モビリティ
吉利（Geely）、フォルクスワーゲン、車載インフラ実装

全二重オムニモーダル通信（MiniCPM-o）
「見ながら、聞きながら、同時に話す」リアルタイム対話、居取り運転番告、調聴アドバイス