

DeepSeek-V4.1-Flash：AIエージェントの経済性を変える新世代モデル

異次元の「低コスト」と「メモリ効率」



競合比20分の1の
圧倒的低価格



\$0.15~
/ 100万トークン

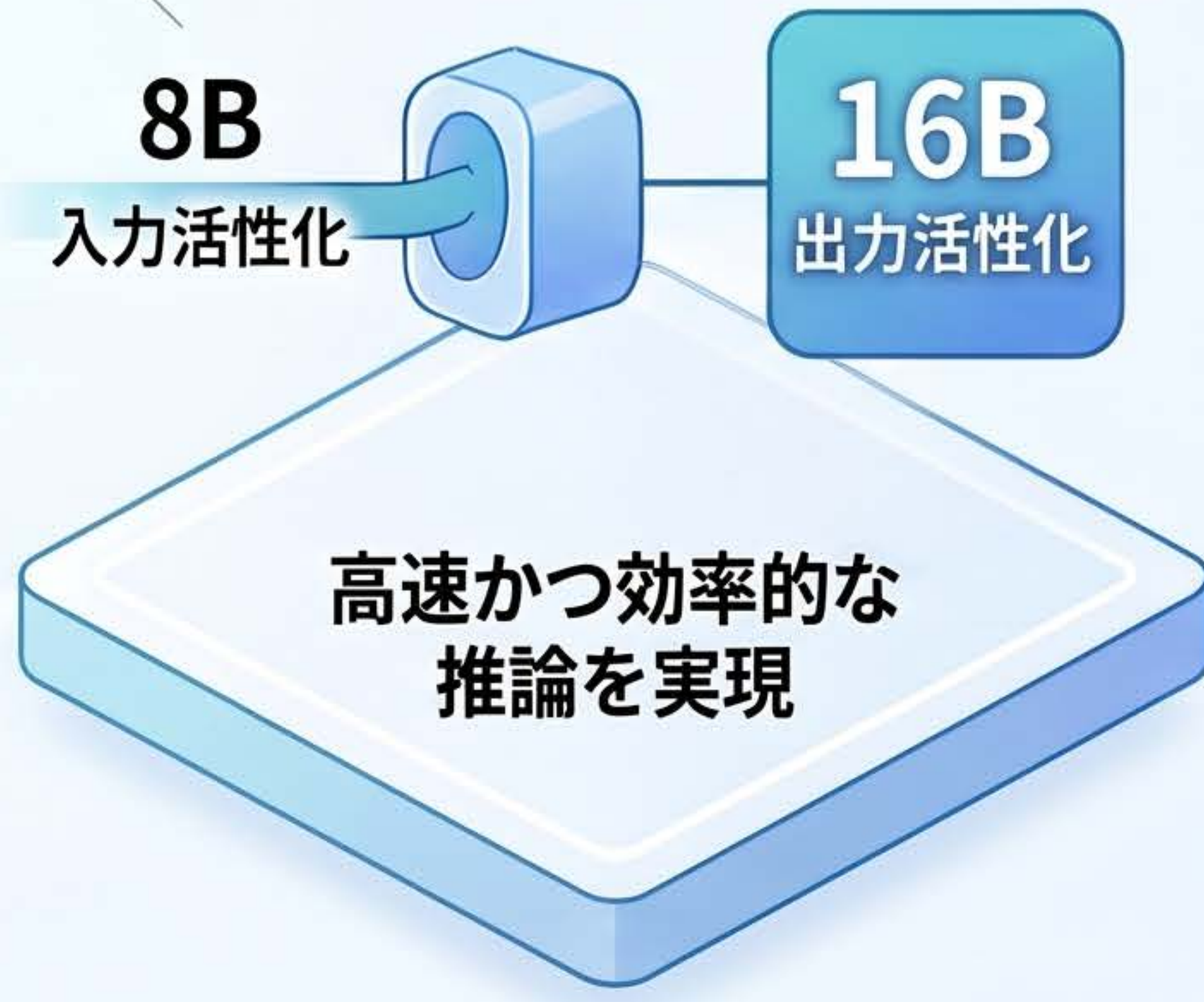
KVキャッシュを
最大437分の1に圧縮



従来

V4.1 Flash

Causal
Encoder-Decoder
(CED)



エージェント性能の実力と導入リスク



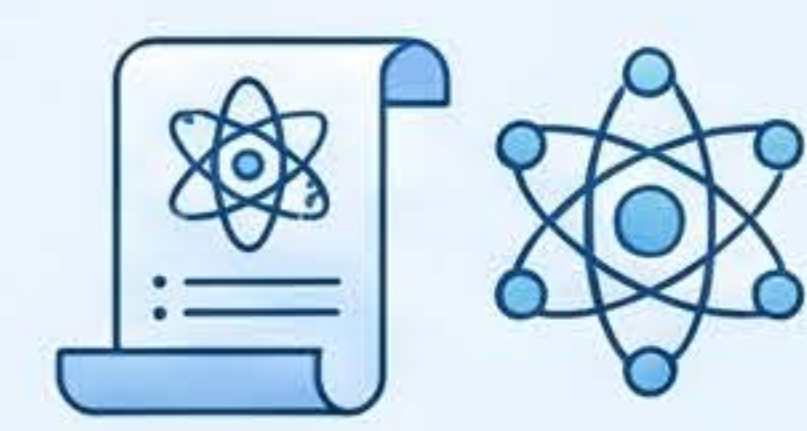
エージェント/コード
開発能力は「Pro」超え

ソフトウェア開発ベンチマークで
旧モデル・競合超え



一般知識と事実性には
「トレードオフ」あり

事実確認・数学では旧Proモデル
に劣り、ハルシネーションに注意



MITライセンスによる
オープンウェイト公開

自社ホスト可能だが、552B級運用
には膨大なGPU資源を要する

モデル名	入力価格 (Cache Miss)	出力価格
DeepSeek-V4.1-Flash	\$0.30	\$1.20
Claude Opus 5	\$5.00	\$25.00
DeepSeek-V4 Pro (旧)	\$1.32	\$3.96

指標 (ベンチマーク)	V4.1 Flash (新)	V4 Pro (旧)	得意・苦手
HumanEval (コード生成)	79.4	76.8	向上
SimpleQA (事実知識)	42.3	55.2	低下
GSM8K (初等数学)	93.0	92.6	同等