

欧州発・主権型AI「Kolibri」の全貌と日本市場へのインパクト

「Kolibri」を支える3つの技術革新

78.1Bの表現力と3.46Bの
アクティブ駆動による高効率化

78.1B

3.46B

総量の約4.4%のみを駆動し、高速推論と低コストを実現。

「Merlin-Arthurプロトコル」による
ハルシネーション抑制



揺らがない場合に回答を累積し、RAGの接地性を向上。

形態素を保持する
独自トークナイザ「UniBPE」

形態素

意味単位を保つトークン分割で処理コストを削減。

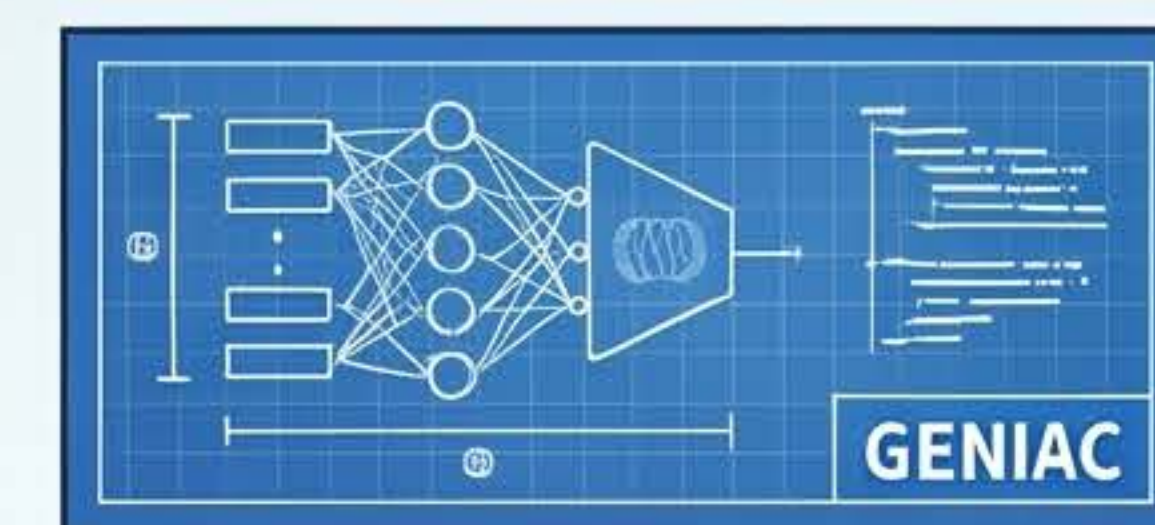
日本のAIエコシステム・政策への影響

富士通「Takane」等を通じた
国内基盤への技術還元



Cohere競合により、叡樹の先端技術が国内モデルへ還元。

国産LLM開発(GENIAC等)における
高効率アーキテクチャの指針



限られた計算インフラでコストパフォーマンスを最大化。

安全保障・重要インフラに向けた
「開かれたAI主権」のモデルケース



エアギャップ運用とデータ退却で信頼性の高い継続基準に。

Kolibri主要スペックと実行環境の要件

項目	技術仕様・構成値
アーキテクチャ	50層 確結合MoE (384エキスパート/層)
パラメータ数	総量 78.1B / トークンあたり有効 3.46B
ライセンス / 運用環境	Apache 2.0 (完全オフライン・エアギャップ対応)