PolySphere-3

AI inside株式会社

モーダル：画像・言語

弊社のPolySphere-3は、他社モデルと比較しても非定型帳票の読み取りの精度が高いことが強みです。このモデルは、弊社独自の読み取りモデルと項目抽出モデルを活用した複合的なモデルです。例えば、製造業の暗黙知においては、手書きメモや検査チェックシートなどの読み取りを行い、これらをデジタルテキスト化することなどが期待できます。

提供形態：自社のＡＰＩを利用した提供
モデル URL：
ユースケース
費用：
問い合わせ先：x-research@inside.ai

Shisa-v2

Shisa株式会社

モーダル：言語

Shisa-v2シリーズは国産の大規模バイリンガルモデルとして、日本語処理で海外モデルを上回る性能を実現。高品質な合成データの活用により、日本語でのタスクで優れたパフォーマンスを示しています。なお、Shisa-v2の405Bモデルは、2025年6月6日現在、国内でトレーニングされたモデルの中で最高のベンチマークスコアを達成しています。

提供形態：モデル公開
モデル URL： https://huggingface.co/collections/shisa-ai/shisa-v2-67fc98ecaf940ad6c49f5689
問い合わせ先： adam@shisa.ai

NABLA-VL

NABLAS

モーダル：言語・画像・動画

画像・動画を入力可能。英語と日本語のパフォーマンスを両立させることを意識した。学習 (ファインチューニング) 用のコードも公開済み。また、トークン削減手法を利用可能にしており、学習時間を約2倍短縮、推論時間を約1.3倍短縮できることを確認済み。

提供形態：モデル公開
モデル URL：
(HF) https://huggingface.co/nablasinc/NABLA-VL,
(GitHub) https://github.com/nablas-inc/NABLA-VL
問い合わせ先：shintate@nablas.com

llm-jp-3.1-1.8b

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3.1-1.8b
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3.1-1.8b-instruct4

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3.1-1.8b-instruct4
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3.1-13b

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3.1-13b
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3.1-13b-instruct4

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3.1-8x13b

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3.1-8x13b
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3.1-8x13b-instruct4

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．日本語MT-Benchのスコアで gpt-4-0613 を上回る性能を記録．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3.1-8x13b-instruct4
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3-13b

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3-13b
問い合わせ先：llm-admin@nii.ac.jp

llm-jp-3-8x13b

NII

モーダル：言語

日本国内でフルスクラッチで学習されたモデルであり，学習コーパス内の日本語比率が高いこと．事前学習コーパスや学習コードを公開しており，学習過程の透明性が高いこと．

提供形態：モデル公開
モデルURL：https://huggingface.co/llm-jp/llm-jp-3-8x13b
問い合わせ先：llm-admin@nii.ac.jp

Sarashina2.2

SB Intuitions株式会社

モーダル：言語

フルスクラッチで構築した日本語大規模言語モデルで、少量のパラメータながら極めて高いベンチマーク性能を達成している

提供形態：モデル公開
モデル URL：
https://huggingface.co/collections/sbintuitions/sarashina2-67c65fdab1ff63d92defb47e
問い合わせ先：geniac-inquiry@sbintuitions.co.jp

Sarashina2-Vision

SB Intuitions株式会社

モーダル：言語・画像

ライセンス上問題のない公開データと合成データのみを使って
学習しており、商用、非商用関わらず利用可能。リリース時点で国内VLM
モデルトップのベンチマークスコアを達成。

提供形態：モデル公開
モデルURL：https://huggingface.co/sbintuitions/sarashina2-vision-14b
問い合わせ先：geniac-inquiry@sbintuitions.co.jp

SG4D10B

SyntheticGestalt株式会社

モーダル：化合物

"100億件の化合物情報で学習させた、世界最大の分子情報に特化した基盤モデル。 新薬・新素材の探索向けに、分子の生成や解析のためのAIモデル開発時に利用ができます。 分子の立体構造情報を
利用している点と大量の低分子化合物情報を学習している点に強みがあり、これまで予測が難しかった新規化合物に対する機械学習モデルの適用を改善する可能性を持っています。"

提供形態：プラットフォーム公開 (Amazon Bedrock等)
モデルURL：
問い合わせ先：k.kamiya@syntheticgestalt.com

Heron-NVILA-Lite-15B

Turing株式会社

モーダル：言語・画像

大規模日本語対応モデルとして、計算効率と応答精度を両立。

提供形態：自社のAPIを利用した提供;プラットフォーム公開 (Amazon Bedrock等)・モデル公開
モデルURL：https://huggingface.co/turing-motors/Heron-NVILA-Lite-15B
問い合わせ先：yukiko.kasai@turing-motors.com

Heron-NVILA-Lite-1B

Turing株式会社

モーダル：言語・画像

軽量かつ高効率な日本語対応モデル。日本国内開発による
高い信頼性と応答品質を持っています。

提供形態：モデル公開
モデルURL：https://huggingface.co/turing-motors/Heron-NVILA-Lite-1B
問い合わせ先：yukiko.kasai@turing-motors.com

Heron-NVILA-Lite-2B

Turing株式会社

モーダル：言語・画像

軽量かつ高効率な日本語対応モデル。日本国内開発による
高い信頼性と応答品質を持っています。

提供形態：モデル公開
モデルURL：https://huggingface.co/turing-motors/Heron-NVILA-Lite-2B
問い合わせ先：yukiko.kasai@turing-motors.com

Heron-NVILA-Lite-33B

Turing株式会社

モーダル：言語・画像

日本語高精度対応の大規模モデル。

提供形態：モデル公開
モデルURL：https://huggingface.co/turing-motors/Heron-NVILA-Lite-33B
問い合わせ先：yukiko.kasai@turing-motors.com

Woven-VLLM

Woven by Toyota

モーダル：言語・画像・動画

独自構築した大規模動画像＋言語のデータセットから学習した
Instance-aware Spatial-Temporal LLM (8B) を採用し、
動画像の理解に最適化。映像理解のベンチマークMVBenchでトップ
レベルのスコアを達成

提供形態：プラットフォーム公開 (Amazon Bedrock等)
モデルURL：
問い合わせ先：quan.kong@woven.toyota

KARAKURI LM 8x7B Instruct v0.1

カラクリ株式会社

モーダル：言語

カスタマーサポートのAI Agentのために作られたモデルです。
使用する際にパラメータによって出力を調整することができるという特徴があります。
それによって真実性を高めたり創作性を高めたりすることができます。

提供形態：プラットフォーム公開 (Amazon Bedrock等)・モデル公開
モデルURL：https://huggingface.co/karakuri-ai/karakuri-lm-8x7b-instruct-v0.1
問い合わせ先：t.nakayama@karakuri.ai

KARAKURI VL 32B Thinking 2507 Experimental

カラクリ株式会社

モーダル：言語、画像

日本語環境でのコンピュータユースを念頭に置いて開発された視覚言語モデルです。

提供形態：プラットフォーム公開 (Amazon Bedrock等)・モデル公開
モデルURL：https://huggingface.co/karakuri-ai/karakuri-vl-32b-thinking-2507-exp
問い合わせ先：t.nakayama@karakuri.ai

KARAKURI VL 32B Instruct 2507

カラクリ株式会社

モーダル：言語、画像

日本語環境でのコンピュータユースを念頭に置いて開発された視覚言語モデルです。

提供形態：プラットフォーム公開 (Amazon Bedrock等)・モデル公開
モデルURL：https://huggingface.co/karakuri-ai/karakuri-vl-32b-thinking-2507-exp
問い合わせ先：t.nakayama@karakuri.ai

Stockmark-2-100B-Instruct-beta

ストックマーク株式会社

モーダル：言語

GENIAC第2期にフルスクラッチで開発された日本語を主な対象とした100Bパラメータの事後学習LLMです。日本語の理解能力に優れ、日本語の
対話能力を測るベンチマークである日本語MT-benchでは、日本でフルスクラッチで開発されたモデルの中では最高性能を示す (2025.03.06公開時点)。また
日本語のビジネスや時事問題のベンチマークではGPT-4oよりも高い性能を示す。

提供形態：モデル公開
モデルURL：https://huggingface.co/stockmark/Stockmark-2-100B-Instruct-beta
問い合わせ先：takahiro.omi@stockmark.co.jp

Stockmark-2-VL-100B-Instruct-beta

ストックマーク株式会社

モーダル：画像

GENIAC第2期にフルスクラッチで開発された主に日本語のドキュメント解析を対象とした100BパラメータのVLMです。日本語のドキュメント理解の性能が高く、スライドの読解のベンチマークであるSlideQAや図表読解のベンチマークであるChartQAではgpt-4oを超える性能を示す。一般の画像理解のベンチマークに
おいても、純国産モデルの中では高い性能を示している。

提供形態：モデル公開
モデルURL：(公開予定)
https://huggingface.co/stockmark/Stockmark-2-VL-100B-Instruct-beta
問い合わせ先：takahiro.omi@stockmark.co.jp

Llama 3.1 Future Code Ja 8B

フューチャー株式会社

モーダル：言語

日本語とソフトウェア開発に特化した基盤モデル。Llama 3.1 8Bを日本語およびソースコード、ソフトウェア開発に関する合成指示チューニングデータを用いて学習したもの。各テーマの開発時に本基盤モデルを用いることで開発を加速させることができると考えられる。

提供形態：モデル公開
モデルURL：
問い合わせ先：pj-geniac@future.co.jp

ABEJA-Qwen2.5-32b-Japanese-v0.1

株式会社ABEJA

モーダル：言語

2025年1月に公開した日本語の継続事前学習をしたQwen2.5-32Bベースのモデルで、汎用的に日本語能力が向上したモデルです。Reasoningモデルと比べて、思考過程を経ないので、速度が求められる場合はこちらを推奨します。

提供形態：モデル公開
モデルURL：https://huggingface.co/abeja/ABEJA-Qwen2.5-32b-Japanese-v0.1
問い合わせ先：kyo.hattori@abejainc.com

ABEJA-Qwen2.5-7b-Japanese-v0.1

株式会社ABEJA

モーダル：言語

32Bモデルに比べて小型である7Bモデルで、こちらも日本語の学習によりベースとなるQwen2.5よりも全体的に日本語性能があがっています。こちらのモデルは蒸留による学習で精度向上をしています。

提供形態：モデル公開
モデルURL：https://huggingface.co/abeja/ABEJA-Qwen2.5-7b-Japanese-v0.1
問い合わせ先：kyo.hattori@abejainc.com

ABEJA-QwQ32b-Reasoning-Japanese-v1.0

株式会社ABEJA

モーダル：言語

o1/o3やDeepSeek-R1のような思考過程を経てから出力をするReasoningモデルです。全体的な性能が高いだけでなく、論理的な思考が必要なタスクを得意としています。

提供形態：モデル公開
モデルURL：https://huggingface.co/abeja/ABEJA-QwQ32b-Reasoning-Japanese-v1.0
問い合わせ先：kyo.hattori@abejainc.com

CommonArt β

株式会社AI Picasso

モーダル：言語・画像

日本語にネイティブ対応している。推論が軽量である。著作権を侵害する恐れがとても少ない。

提供形態：モデル公開
モデルURL：https://huggingface.co/aipicasso/commonart-beta
問い合わせ先：ozaki@aidealab.com

AIdeaLab VideoJP

株式会社AIdeaLab

モーダル：言語・動画

日本語のネイティブ入力に対応している。生成にかかる時間も短い。著作権を侵害する恐れがとても少ない。

提供形態：モデル公開
モデルURL：https://huggingface.co/aidealab/AIdeaLab-VideoJP
問い合わせ先：ozaki@aidealab.com

ELYZA-Shortcut-1.0-Qwen-32B

株式会社ELYZA

モーダル：言語

Reasoning Model の開発過程で生成されたデータを使用して、「Reasoning Model が深く思考して辿り着いた回答を、反射的に答えられるように暗記したモデル」の学習を行いました。320億パラメータと軽量ながら、同じ非 Reasoning Model で商用の API が展開されている OpenAI 社の「GPT-4o」に匹敵する性能を達成しています。

提供形態：モデル公開
モデルURL：https://huggingface.co/elyza/ELYZA-Shortcut-1.0-Qwen-32B
問い合わせ先：takuya.harada@elyza.ai

ELYZA-Shortcut-1.0-Qwen-7B

株式会社ELYZA

モーダル：言語

Reasoning Model が深く思考して辿り着いた回答を、反射的に答えられるように暗記したモデル (= Shortcut Model) です。ベースモデル (Qwen2.5-7B-Instruct) と比較してJMMLU、Japanese MT-Bench、ELYZA Tasks 100 といった日本語能力に関するベンチマークのスコアが向上しています。

提供形態：モデル公開
モデルURL：https://huggingface.co/elyza/ELYZA-Shortcut-1.0-Qwen-7B
問い合わせ先：takuya.harada@elyza.ai

ELYZA-Thinking-1.0-Qwen-32B

株式会社ELYZA

モーダル：言語

本モデルは OpenAI 社の「o1/o3」シリーズや、DeepSeek 社の「DeepSeek-R1」と同様に、思考の連鎖 (Chain of Thought; CoT) を通して複雑な論理的思考を行う能力を強化した Reasoning Model です。

提供形態：モデル公開
モデルURL：https://huggingface.co/elyza/ELYZA-Thinking-1.0-Qwen-32B
問い合わせ先：takuya.harada@elyza.ai

PLaMo

株式会社Preferred Networks

モーダル：言語

日本語性能において高いベンチマークスコアを示し、第三者評価の結果においても例えば基礎言語能力ではGPT等を抑えて1位になる等、高い評価を得ている。OpenAI完全互換の形式で独自APIやAWS Bedrock経由での利用環境を用意。2Bの小型モデルも用意しており、特定用途に応じてオンプレでの推論環境構築も可能。工場内や官公庁におけるオンプレ実装や、CS業務における自然な日本語表現などが有用と想定。

提供形態：自社のAPIを利用した提供・プラットフォーム公開 (Amazon Bedrock等)・モデル公開
モデルURL：https://plamo.preferredai.jp/api https://aws.amazon.com/marketplace/pp/prodview-eybaqh3shlfmk https://aihub.qualcomm.com/models/plamo_1b https://huggingface.co/pfnet
問い合わせ先：plamo-support@preferred.jp

calm3-22b-chat

株式会社サイバーエージェント

モーダル：言語

商用利用可能なApache2.0ライセンスでオープンモデルとして提供

提供形態：モデル公開
モデルURL：https://huggingface.co/cyberagent/calm3-22b-chat
問い合わせ先：ishigami_ryosuke@cyberagent.co.jp

DATAGRID-Local-Attention-DiT-v1.0.0-0.52B

株式会社データグリッド

モーダル：画像

PixArt-αのDiffusion Transformerをベースに、従来のグローバル注意機構の代わりに独自設計の局所的注意機構 (Local Attention) を導入することで、計算効率の向上とパラメータ数の削減を実現している。

提供形態：モデル公開
モデルURL：https://huggingface.co/DATAGRID-research/DATAGRID-Local-Attention-DiT-v1.0.0-0.52B
問い合わせ先：yu.saito@datagrid.co.jp

DATAGRID-Open-Sora-Plan-v1.3.0-0.16M

株式会社データグリッド

モーダル：動画

Open-Sora-Plan_v1.3を基盤として、ロイヤリティフリーの
動画共有サイトから収集した16万の動画データセットをベースにクレンジングしたデータセットで学習したモデルで、学習可能なオープンモデルでFVD (sky_timelapse) においてSOTAを記録

提供形態：モデル公開
モデルURL：https://huggingface.co/DATAGRID-research/DATAGRID-Open-Sora-Plan-v1.3.0-0.16M
問い合わせ先：yu.saito@datagrid.co.jp

DATAGRID-stable-diffusion-inpainting-manufacturing-fulltuned

株式会社データグリッド

モーダル：画像

pixabayなどの画像共有サイトから様々な製品画像 (金属表面画像等) を収集し、構築した製造業特化のインペインティング型の画像生成モデルで、製造業における外観検査の不良品画像生成に使用することができます。物体転写の前景に関するLPIPS, CLIP, DISTSなどの指標でSOTAを達成しています。

提供形態：自社のAPIを利用した提供
モデルURL：
問い合わせ先：yu.saito@datagrid.co.jp

Ubitus Multi-language Llama 3.1

株式会社ユビタス

モーダル：言語・画像・音声

本モデルはMeta Llama3.1 405Bに事前追加学習とファインチューニングを施した日本語最適化LLM。日本語タスク実行に優れた命令チューニングとDPOにより大幅な性能向上を実現。各種日本語ベンチマークではGPT-4 Turboを上回り、Llama 3.1インストラクトと比較して日本語ベンチマークが11%向上。日本語の推論、対話、理解タスクに優れ、最高精度の多言語スケーリングと再現性を提供。
提供形態：自社のAPIを利用した提供・プラットフォーム公開 (Amazon Bedrock等)・モデル公開
モデルURL：
問い合わせ先：contact@ubitus.net

Llama-3.1-70B-Instruct-multimodal-JP-Graph-v0.1

株式会社リコー

モーダル：言語・画像

日本のドキュメントに特有な複雑な図表の読み取り精度が向上し、同様の規模のオープンソースを凌ぐ性能を持つ (JDocQAにて測定)

提供形態：モデル公開
モデルURL：
問い合わせ先：https://promo.digital.ricoh.com/ai/contact/

Llama 3.3 Swallow 70B Climate-Instruct v0.1

国立研究開発法人海洋研究開発機構

モーダル：言語

気候変動の学術知識を有し、国内の温暖化対策立案に特化した唯一のモデル。自治体等における温暖化対策のチェックや生成業務の効率化に活用可能。

提供形態：リクエストに応じて提供
モデルURL：
問い合わせ先：daisuke@jamstec.go.jp

Llama 3.3 Swallow 70B Instruct v0.4

東京科学大学

モーダル：言語

Llama 3.3 Swallow 70B Instruct v0.5はLlama 3.3をベースに日本語の能力を強化した大規模言語モデルです。日本語理解・生成タスクにおいて、Qwen2.5-72Bとほぼ同等で、GPT-4oに迫る性能を達成しています。Llama 3.3ライセンスに従い、かつGemma利用規約の利用制限に抵触しない範囲で、研究や商業目的などで利用できます。

提供形態：モデル公開
モデルURL：https://huggingface.co/tokyotech-llm/Llama-3.3-Swallow-70B-Instruct-v0.4
問い合わせ先：okazaki@comp.isct.ac.jp

Gemma-2-Llama Swallow 27B IT v0.1

東京科学大学・産業技術総合研究所

モーダル：言語

Gemma-2-Llama Swallow 27B IT v0.1はGemma 2をベースに日本語の能力を強化した大規模言語モデルです。日本語の理解・生成・対話タスクにおいて、27B以下の規模のLLMの中でトップクラス、70B規模のLLMに迫る性能を達成しています。Gemma利用規約の利用制限に抵触せず、かつLlama 3.3ライセンスに従う場合において、研究や商業目的等で利用できます。

提供形態：モデル公開
モデルURL：https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-27b-it-v0.1
問い合わせ先：okazaki@comp.isct.ac.jp

Gemma-2-Llama Swallow 2B IT v0.1

東京科学大学・産業技術総合研究所

モーダル：言語

Gemma-2-Llama Swallow 2B IT v0.1はGemma 2をベースに日本語の能力を強化した大規模言語モデル (LLM) です。日本語の理解・生成・対話タスクにおいて、2B以下の規模のLLMの中でトップクラス、7B規模のLLMに迫る性能を達成しています。Gemma利用規約の利用制限に抵触せず、かつLlama 3.3ライセンスに従う場合において、研究や商業目的などで利用できます。

提供形態：モデル公開
モデルURL：https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-2b-it-v0.1
問い合わせ先：okazaki@comp.isct.ac.jp

Gemma-2-Llama Swallow 9B IT v0.1

東京科学大学・産業技術総合研究所

モーダル：言語

Gemma-2-Llama Swallow 9B IT v0.1はGemma 2をベースに日本語の能力を強化した大規模言語モデル (LLM) です。日本語の理解・生成・対話タスクにおいて、9B以下の規模のLLMの中でトップクラス、27B規模のLLMに迫る性能を達成しています。Gemma利用規約の利用制限に抵触せず、かつLlama 3.3ライセンスに従う場合において、研究や商業目的等で利用できます。

提供形態：モデル公開
モデルURL：https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1
問い合わせ先：okazaki@comp.isct.ac.jp

Llama 3.1 Swallow 8B Instruct v0.3

東京科学大学・産業技術総合研究所

モーダル：言語

Llama 3.1 Swallow 8B Instruct v0.3はLlama 3.1の英語の能力を維持しながら、日本語の能力を強化した大規模言語モデルです。コンパクトではありますが、日本語の理解・生成・対話タスクで高い性能を達成しています。Llama 3.1ライセンスに従い、かつGemma利用規約の利用制限に抵触しない範囲で、研究や商業目的などで利用できます。

提供形態：モデル公開
モデルURL：https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3
問い合わせ先：okazaki@comp.isct.ac.jp

Fujitsu-LLM-KG

富士通株式会社

モーダル：言語

46.7Bパラメータ のLLMをナレッジグラフで論理推論に特化させ、GPT-4を凌ぐ日本語QA精度を達成。継続事前学習方式と推論手順を公開。また、同様の手法で文書レベル関係抽出精度も向上、GPT-4単体を上回り、日本語で世界最高レベルの精度を実現。

提供形態：モデル公開
モデルURL：https://huggingface.co/Fujitsu-LLM-KG
問い合わせ先：fj-GENIAC-fjgroup@dl.jp.fujitsu.com

CellScribe

株式会社ヒューマノーム研究所

モーダル：遺伝子発現量データ

CellScribeは、約3億細胞もの大規模かつ高品質な遺伝子発現データで学習した基盤モデルです。低品質データの除去、メタデータの精緻化、新規データベースの追加により、先行モデルから質・量ともに大幅に強化されています。強化された前処理機能やデータ拡張機能といった独自のアーキテクチャ改良により、と比較し、遺伝子発現量予測において 7.8%の予測誤差低減という性能優位性を示しています。

提供形態：モデル公開
モデルURL：https://bit.ly/CellScribe
問い合わせ先：info@humanome.jp

tsuzumi

株式会社NTTデータ

モーダル：言語

tsuzumiはNTTが開発する大規模言語モデルです。NTTが長年培ってきた自然言語処理技術により、軽量ながらも高性能なモデルを実現しています。特に日本語の処理能力において優れた性能を発揮します。特定の業界・業務に対してチューニングすることで幅広い分野で活用いただけます。

提供形態：個別対応
問い合わせ先：tsuzumi-nttdata@hml.nttdata.co.jp

VoiceCore

有限会社GIPU

モーダル：入力された日本語テキストを音声に変換

日本語のテキストを感情豊かな日本語音声に変換するVoice AIエージェントです。AIが人間と音声を使った意思疎通を行う事を容易にするために設計されており、非言語音声や感情表現が可能であることが特徴です。

提供形態：モデル公開
モデルURL：https://huggingface.co/webbigdata/VoiceCore
問い合わせ先：gipu@webbigdata.jp

Sparticle/DeepSeek-R1-Distill-Llama-70B-AWQ

Sparticle株式会社

モーダル：言語

ハイブリッド量子化技術により、精度をほぼ維持したままGPU使用量を約85%削減、推論速度を143%～159%向上させ、H100×2枚の環境で効率的に稼働できる点が特徴です。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/Llama-3.3-70B-Instruct_gptq_gs_32

Sparticle株式会社

モーダル：言語

混合精度量子化：メモリ使用をほぼ4ビット並みに抑えつつ、一部を高精度に保持することで精度を大幅向上。実用性が高い。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/llama3.1-70B-4bit

Sparticle株式会社

モーダル：言語

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルにもスケーラブル。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/R1-Distill-Llama-70B_gptq_gsize_32

Sparticle株式会社

モーダル：言語

混合精度量子化：メモリ使用をほぼ4ビット並みに抑えつつ、一部を高精度に保持することで精度を大幅向上。実用性が高い。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/llama3.3-70b-ins-gptqint4

Sparticle株式会社

モーダル：言語

事後量子化手法の中でも精度劣化が少なく、特に学習済みLLMへの適用が容易。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/llama3.1-405B-awq-4bit

Sparticle株式会社

モーダル：言語

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルにもスケーラブル。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Qwen/Qwen2-VL-72B-Instruct-AWQ

Sparticle株式会社

モーダル：言語、動画

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルにもスケーラブル。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Qwen/Qwen2.5-VL-72B-Instruct-AWQ

Sparticle株式会社

モーダル：言語、動画

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルにもスケーラブル。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/DeepSeek-R1-Distill-Llama-70B-AWQ

Sparticle株式会社

モーダル：言語

グリッド量子化技術により、精度をほぼ維持したままGPU使用量を105%削減、推論速度も143% ~ 159%向上させ、1H00×2の効率と信頼性で活用できる点が特徴です。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/Llama-3.3-70B-Instruct_gptq_gs_32

Sparticle株式会社

モーダル：言語

混合精度量子化 メモリ使用がほぼ2ビット並みに抑えつつ、一倍を高精度に保持することで精度を大幅向上。実用性が高い。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/Llama3_3_70B-ins-gptqint4

Sparticle株式会社

モーダル：言語

事後量子化手法の中でも精度劣化が少なく、特に学習済みLLMへの適用が容易。

提供形態：自社のＡＰＩを利用した提供;GBase On-premisesソリューションとして
問い合わせ先：tangjs@sparticle.com

Sparticle/llama3.1-405B-awq4bit

Sparticle株式会社

モーダル：言語

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルもスケールアップ。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Qwen/Qwen2-VL-72B-Instruct-AWQ

Sparticle株式会社

モーダル：言語、画像

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルもスケールアップ。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/Llama-3.1-70B_gptq_gs_32

Sparticle株式会社

モーダル：言語

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルもスケールアップ。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Sparticle/Llama-3.3-70B-Instruct_gptq_gs_32

Sparticle株式会社

モーダル：言語

混合精度量子化 メモリ使用がほぼ2ビット並みに抑えつつ、一倍を高精度に保持することで精度を大幅向上。実用性が高い。

提供形態：GBase On-premisesソリューションとしての提供;
問い合わせ先：tangjs@sparticle.com

Qwen/Qwen2-VL-72B-Instruct-AWQ

Sparticle株式会社

モーダル：言語、画像

前処理不要で高精度・高速な推論を実現する量子化手法。大規模モデルもスケールアップ。

提供形態：GBase On-premisesソリューションとしての提供、
プラットフォーム公開（Amazon Bedrock等）
問い合わせ先：tangjs@sparticle.com

Fast-Math-Qwen3-14B
アイリス株式会社

モーダル：言語

Qwen3に対して追加の教師あり学習と強化学習を実施し、数学推論能力と推論効率（答えに達するまでの出力トークン数）を改善したモデル。オリジナルのQwen3モデルと比較し、数学タスクにおける精度を維持しつつも推論速度を最大で65%高速化。Kaggleコンペティション「AI Mathematical Olympiad - Progress Prize 2」にて金メダルを獲得。

提供形態：モデル公開（GitHub等）
問い合わせ先：tangjs@sparticle.com
hiroshi.yoshihara@aillis.jp

MedExamDoc-Llama-3.1-Swallow-8B-Instruct-v0.5
インジェンタ株式会社

モーダル：言語

多言語対応しつつも海外モデルと違い国内利用に主軸を置き、創薬医療に特化した日本語のデータでチューニング済。再生医療企業の業界知見をもとにを、日本語で有機的に構造化し（知識Graph）、既存のデータベースに含まれる遺伝子・分子それぞれ独立した情報が医療情報との有機的な連携を実現している。特化型ゆえの業界前提知識でR＆Dや顧客対応において自社内の蓄積をこれまでにないスピードと精度で活用できる。

提供形態：モデル公開（GitHub等）
問い合わせ先：c.lin@ingenta.ai　080-2436-1450

RakutenAI-7B
楽天グループ株式会社

モーダル：言語

高品質な日本語・英語データで事前学習されており、日本語に最適化した独自の形態素解析器を採用。リリース時点において、オープンな日本語LLMの中でトップクラスの評価を獲得しており、自然で対話型の会話に優れている。指示の正確な理解と効率的なリソース使用を両立し、限られたリソースでも多岐にわたるタスクにおいて高いパフォーマンスを発揮。

提供形態：モデル公開（GitHub等）
問い合わせ先：aidd-geniac@rakuten.com

RakutenAI-2.0-8x7B
楽天グループ株式会社

モーダル：言語

RakutenAI-7Bの能力を基盤とし、さらに高度化した大規模モデル。8つのエキスパートが連携するMixture of Experts (MoE) アーキテクチャを採用することで、複雑なタスクにおいてより優れたパフォーマンスを実現。膨大な知識と文脈を正確に把握し、創造的かつ論理的な応答を生成可能。企業利用、研究、その他多様な分野において革新的なソリューションを提供

提供形態：モデル公開（GitHub等）
問い合わせ先：aidd-geniac@rakuten.com

RakutenAI-2.0-mini
楽天グループ株式会社

モーダル：言語

RakutenAI-2.0-8x7Bの高性能を維持しつつ、軽量化と高速応答を追求したコンパクトモデル。エッジデバイスやモバイル環境での利用に最適化されており、限られた計算資源でも高い精度と迅速な処理を実現。組み込みシステムやリアルタイム対話アプリケーションなど、多様なユースケースでAIの恩恵を広げ、ユーザー体験を向上。

提供形態：モデル公開（GitHub等）
問い合わせ先：aidd-geniac@rakuten.com

ARUMCODE 1 -5axis
大矢伝動精機株式会社

モーダル：言語;ジオメトリー解析・テキスト

製造ＡＩ「ＡＲＵＭＣＯＤＥ」は世界初の人工知能であり、ＡＲＵＭＣＯＤＥを搭載したＴＴＭＣも類似製品はありません。アルムはソフトウェア検証、実加工検証を自社実施してＡＩ開発環境・体制を構築している。弊社として切削加工における完全自動化とＡＩ化を検証する上でＡＲＵＭＣＯＤＥとＴＴＭＣを稼働することは、アルム製品を米国・インド・韓国に輸出すること、「キーパーツ調達プラットフォーム事業」に不可欠である

提供形態：ＡＲＵＭＣＯＤＥはサブスクリプション（Microsoft Azure）または既存工作機械に標準搭載（産業用PCにインストール）した形での提供。ＴＴＭＣは売り切り販売
問い合わせ先：523313141