特化型モデル開発のためのモデルの小型化

実施者	株式会社ABEJA
事業概要	特定タスクにおいて高性能で実用可能でありながら、モデルのパラメータサイズの規模を抑えた実運用を見据えたLLMを開発する。

精度とコストのトレードオフ

LLMは大きくなるほど計算資源とコストが増え、実運用では精度とコストのトレードオフが課題となる。
その課題解決に向けて、実運用上必要な特定タスク性能をあげつつも、小型化したLLMを開発する。

社会実装イメージ

●ABEJA Platformへの搭載
●ABEJA Platformを通じて各顧客企業へ導入
●ABEJA Platformを基盤とした各顧客企業業務のLLM連携による生産性向上
●エッジ環境へ実装し、活用範囲の拡大

事業成果

①3つのLLMの構築

モデル	性能
ABEJA Qwen2.5-32B Model	Open AI GPT-4超え (2025/01時点)
ABEJA QwQ-32B Reasoning Model	OpenAI GPT-4oおよびo1-preview超え(2025/04時点)
ABEJA Qwen2.5-7B Model	同規模モデルの最高水準かつOpenAI GPT-3.5 Turbo超え(2025/04時点)

▶ ABEJA QwQ-32B Reasoning Modelにて目標としたMT-Bench Japanese 8.278を上回る8.669を達成した。

加えて、他の2モデルに先駆けて構築し公開した ABEJA Qwen2.5-32B Modelは、社外でも高評価を獲得（2025年1月公開）した。

- 日本経済新聞社「NIKKEI Digital Governance」「AIモデルのスコア化ランキング」（25年3月公開）において、日経首位、グローバル16位
- Wandb社のNejumi Leaderboard3の総合スコアでは50B以下のモデルとして1位を獲得

②情報の公開

モデル：3つのモデル全てをHuggingfaceで公開済み
開発ノウハウ：テックブログ及び登壇にて展開済み