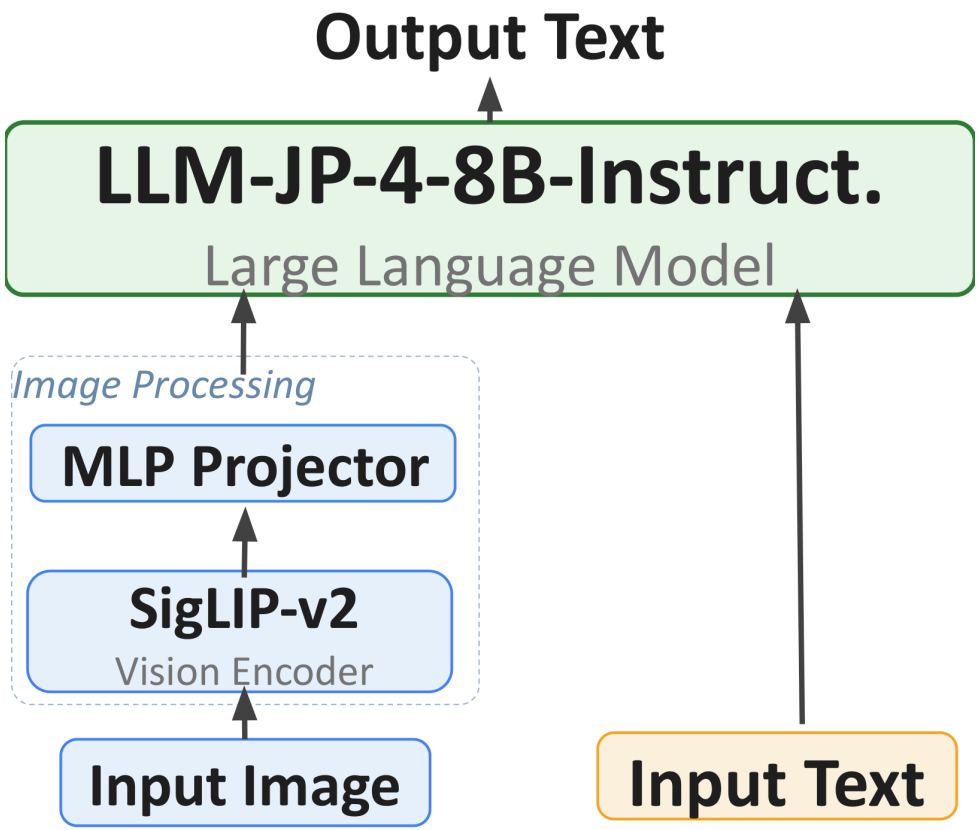
推論強化と文書読解の統合による日本語金融VLMの開発

COMPASS — FT-LLM 2026

柳澤 篤 ・ 柿本 源心
京都大学 — 京大OLC

概要・モデルアーキテクチャ

- 日本語の金融文書を**画像から直接理解・推論**できるVLMを開発
- 官公庁PDFと数学データから**独自データセットを構築**
- **8Bパラメータ**で大規模モデルに迫る性能を実現
- 3段階の学習パイプラインで段階的にファインチューニング



学習パイプライン（Phase 1 / 2 / 3）

Phase 1 VLM基盤構築

- Stage 1: キャプション・OCRデータで**Projectorのみ学習**
- Stage 2: VQAデータで**Encoder + Projector + LLM（LoRA）を全更新**

Phase 2 推論能力の強化

- Qwen3-30Bからの**知識蒸留**（数学データ約10万件）
- SFT で**LLMのみ学習**
- GSM-8K: 64.6 → **73.2**（+8.6pt）

Phase 3 金融ドメイン特化

- Stage 1: 既存金融QA 3種で**LLMのみ学習**
- Stage 2: 官公庁PDFから**画像QA約3.2万件**を作成し、Projector + LLM（LoRA）を更新

結果とまとめ

主要な評価結果

ベンチマーク	Phase 1	Phase 2	Phase 3
GSM-8K (Acc.%)	64.6	**73.2**	71.9
不正検知 (AUC)	0.472	0.534	**0.580**
JP Fin Harness (Acc.%)	54.2	31.6	49.6

- Phase 2で**数学推論が大幅向上**（GSM-8K +8.6pt）
- Phase 3で**金融タスク性能を回復**
- 不正検知AUC 0.580は**70B Llama（0.59）に匹敵**

結論

- 官公庁PDFからのデータ構築と3段階パイプラインで**8Bの日本語金融VLM**を実現
- 適切なデータキュレーションと段階的学習で、**大規模モデルに迫る金融分析能力**を達成

課題・Limitation

- Phase 2で金融タスクの性能が一時低下し、Phase 3で回復（タスク依存）
- VLMとしての画像QA精度にはまだ改善の余地あり