

# IIIF Tsukushi Viewer : 生成AIチャット がアシストする日本古典籍の新しい 読み方



北本 朝展 (KITAMOTO Asanobu)

ROIS-DS人文学オープンデータ共同利用センター (CODH)

国立情報学研究所

**共同研究者** : 本間 淳 (フェリックス・スタイル) 、カ  
ラーヌワット・タリン (Sakana AI)



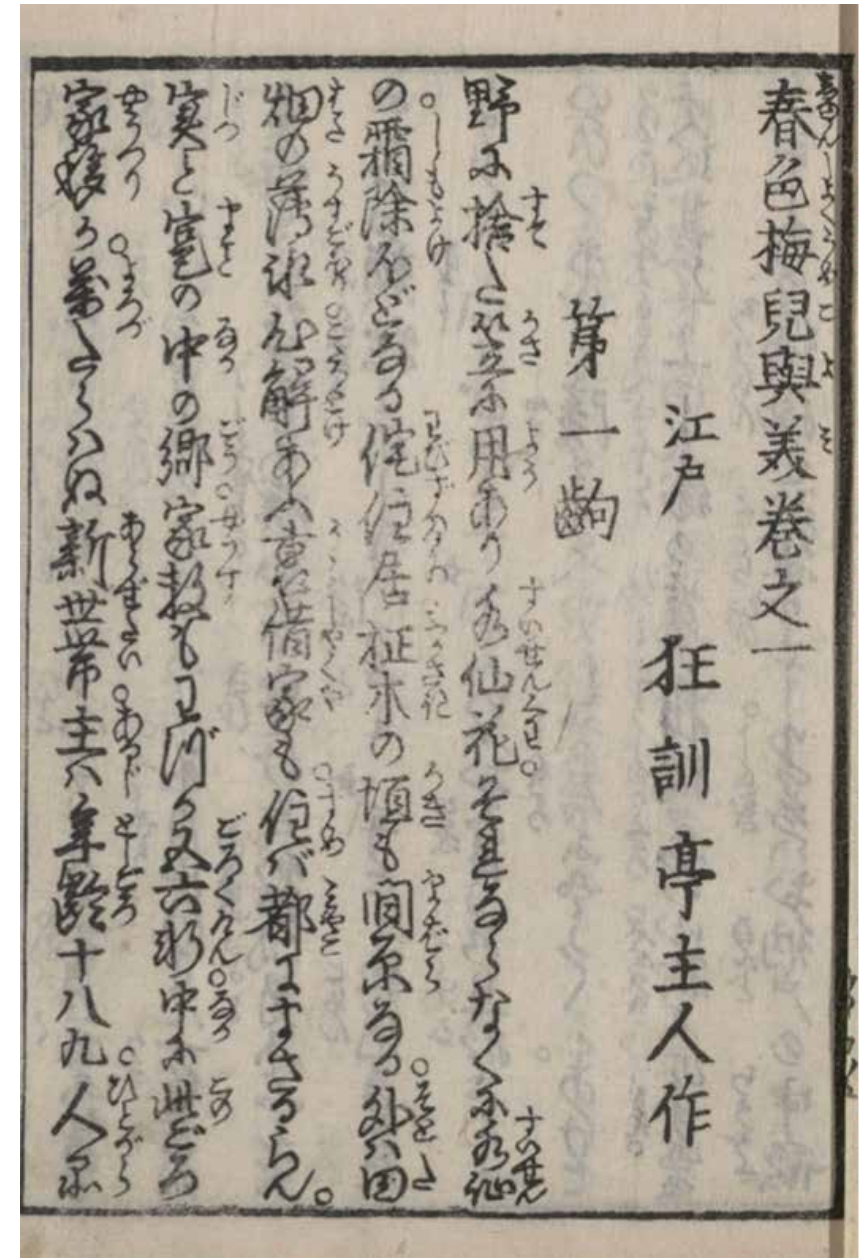
# 「つくし」プロジェクト

<http://codh.rois.ac.jp/tsukushi/>

1. 課題：LLMの学習データに日本語テキスト（特に歴史的日本語）が不足（low-resource language）
2. AIくずし字OCRによる大規模データセット構築
3. 大規模言語モデルを活用するためのアプリ開発

# AIアシストで古典籍を読む

1. **文字の問題**：AIくずし字認識により解決
2. **文法・語彙の問題**：生成AIチャットにより解決
3. **表示の問題**：IIF (International Image Interoperability Framework) により解決
4. **古典籍の新しい読み方を提案**



絵本江戸桜

前 4 / 16 次

☆ 画像 ダウンロード 共有 ヘルプ



AIチャット

チャット対象

本全体 ページリスト 画像領域

あなた  
本全体の内容をまとめて

AI (gpt-4o)  
「絵本江戸桜」は江戸の名所や風景を描いた絵本で、寛政7年正月に序文が書かれています。内容は、江戸桜の美しさや、浅草寺、寛永寺といった江戸の主要な寺社・名所が紹介されています。浅草寺の本尊である観音菩薩や、寛永寺の薬師如来など宗教的な要素も含まれています。また、新吉原の遊郭や両国橋の夏の納涼風景なども詳細に描写されています。全体として、江戸の風景や文化の豊かさが強調されています。

チャットを続ける

要約する

IIIF画像  
ビューア

生成AI  
チャット

# 生成AIチャット

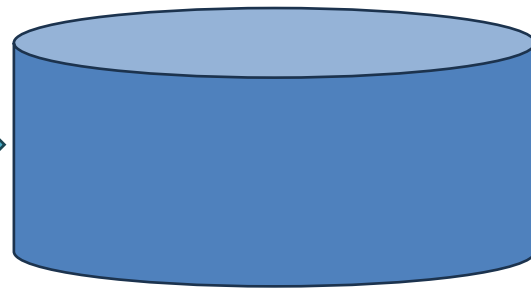
プロンプト

この本の名所ガイド  
を作って

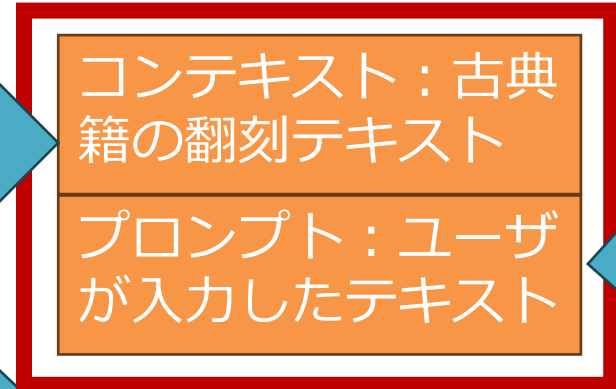


くずし字OCR

IIIF識別子の  
リスト



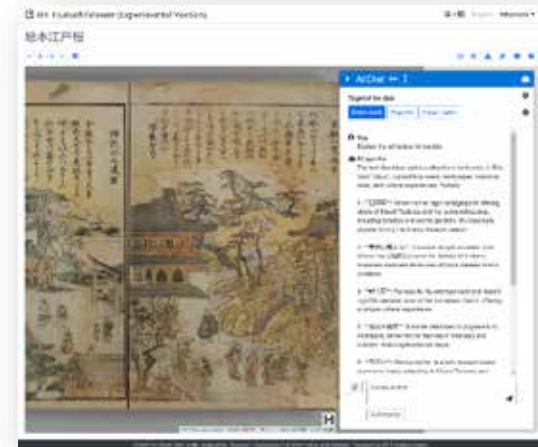
Content Service



くずし字OCRしたテキ  
スト（IIIF識別子単位でア  
クセス可能）

LLM

Chat  
Service



# 生成AIチャットのコンポーネント

1. **Tsukushi Content Service** : IIFの基本単位でテキストを登録し検索できる、RESTful APIのウェブサービス
2. **Tsukushi Chat Service** : 利用者から受け取ったプロンプトを大規模言語モデルに伝え、その回答を利用者に届ける、JSON-RPC APIのウェブサービス
3. **IIF Tsukushi Viewer** : 生成AIチャットの機能を備えたIIF画像ビューア

# 本の情報単位（粒度）とIIIFの関係

本の情報単位（粒度）	IIIFの単位と識別子
本全体	Manifest (Manifest URI)
ページリスト	Canvas (Canvas ID)
画像領域	Canvasの一部 (Image Request URI)

# Tsukushi Content Service

1. Manifest URIとCanvas IDをキーとして、くずし字OCR (RURI) の結果を登録し、API経由で提供
2. Manifest URIをキーとして、本に関するメタデータを登録し、API経由で提供
3. 1に対して、テキスト埋め込みモデル (OpenAI text-embedding-3-small) によるベクトル化を実行
4. PostgreSQLとpgvectorにより、SQLを用いた柔軟な検索と、高速なベクトル検索を統合



# 課題：本に対する生成AIチャット

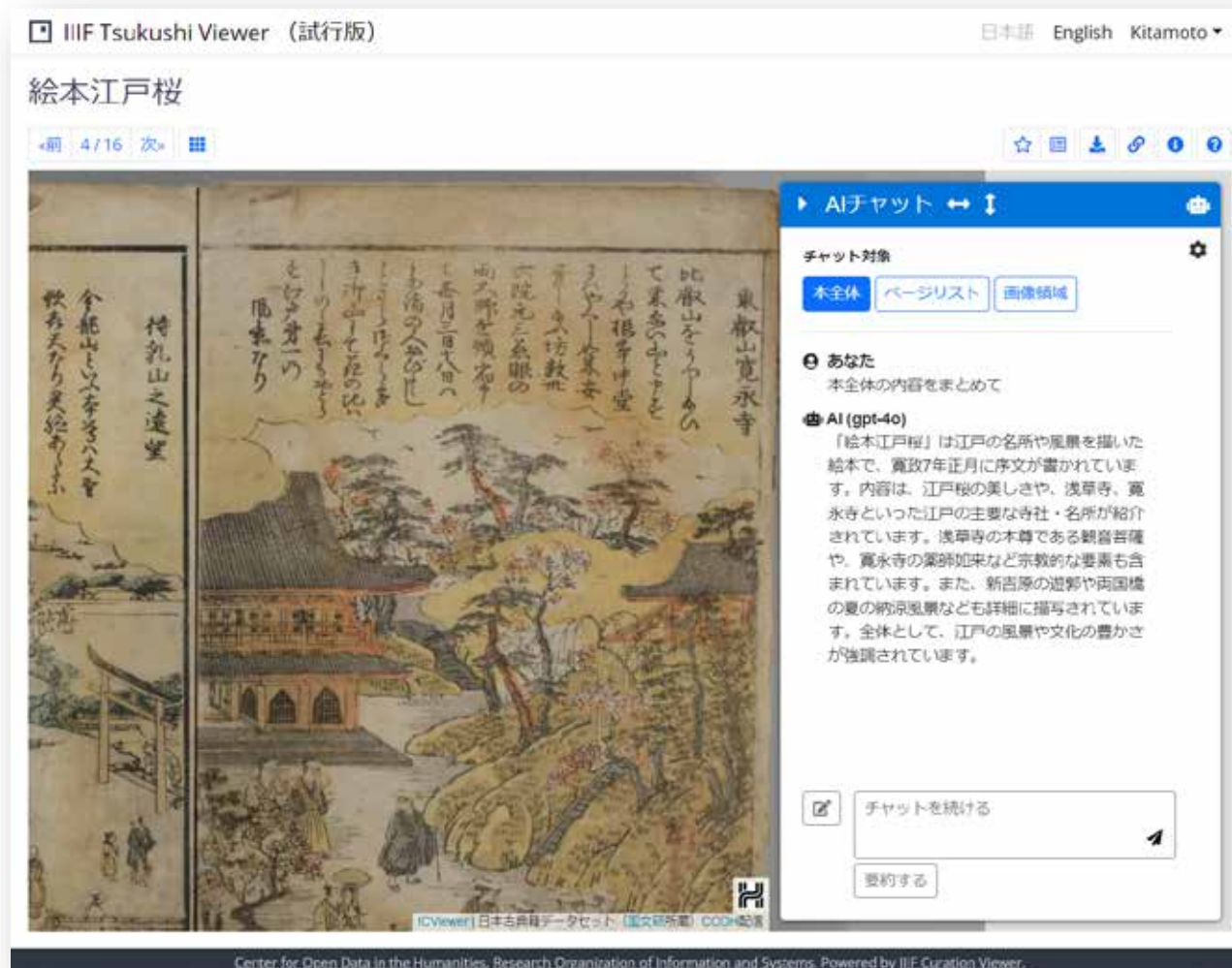
1. プロンプトのテンプレートを「**〇〇に対して××せよ**」（**対象選択→指示実行**）と定義
2. **プロンプトの例**：「本全体に対して要約せよ」「ペー  
ジリストに対して翻訳せよ」
3. **チャットの対象**→本に対する情報単位（粒度）を決定
4. **チャットの指示**→自由記述に加えて、「要約する」な  
どの定番プロンプトをライブラリ化

# Tsukushi Chat Service

1. **JSON-RPC形式のAPI**として、**多種のメソッドを単一のエンドポイントに実装**
2. **メソッド**：対象と指示タイプをつなげたメソッド名を定義（「book:chat」「text:summarize」など）
3. **パラメータ**：各メソッドに対するパラメータを、値またはJSON形式で指定
4. **ID**：「同一IDなら一連のチャット」と定義することで、チャットのコンテキスト切り替えを制御

# IIF Tsukushi Viewer

<http://codh.rois.ac.jp/software/iif-tsukushi-viewer/>



1. 継承：IIF Curation Viewerの基本機能
2. 改良：キュレーション機能をページリスト機能に転用
3. 追加：2つのAPIと連携して動作するAIチャット機能

# 現代日本人と古典籍との距離を縮めるAI

<http://codh.rois.ac.jp/tsukushi/search/>



1. つくしサーチ：現代日本語で、ページを検索
2. プロンプト：現代日本語で、指示を入力
3. チャット：現代日本語で、AIと対話
4. 現代日本語だけで、古典籍と対話できる

# チャットの例：検索拡張生成（RAG）



1. 入力されたプロンプトをベクトル化
2. ベクトル検索を用いて、類似ページのテキストを取得
3. 取得テキストをコンテキストに設定し、回答を生成

# LLM + 検索か、検索 + LLMか？

1. 検索拡張生成（RAG）はLLMありき = LLMにない知識を外から補うために、**LLMに検索を後付け**
2. 資料を読みながらLLMを使う = 資料の内容を理解することが目的であり、**LLMだけで完結させる意味がない**
3. **検索の入力と出力にLLMを後付けする**、生成拡張検索（GAR）として設計する方が自然
4. **知識と言語の分離**：LLMに期待する役割は、知識の提供よりも、情報の変換（のための言語能力）

# 持続可能なサービス設計

1. 価値は高いが、コストも高いサービスをどう使うか？
2. 一律の利用回数制限は、ユーザ獲得の障害に
3. ローカルLLMも性能・計算資源のスケールングに課題
4. APIサービスの価格下落により、API利用の優位性が増大？

The screenshot shows the 'AIチャット設定' (AI Chat Settings) interface. Under 'モデル設定' (Model Settings), there are buttons for '高性能優先' (High Performance Priority), '利用回数優先' (Usage Count Priority), and 'すべて' (All). Below that, there are buttons for '多言語' (Multilingual), '日本語中心' (Japanese Centered), '英語中心' (English Centered), and 'すべて' (All). A grid of model buttons is shown, including 'claude-3-5-sonnet☆☆☆', 'gpt-4o☆☆☆', 'gemma2☆☆', 'qwen2☆☆', 'elyza☆☆', 'phi-3☆', 'llm-jp☆', 'mistral☆', 'llama3☆', and 'aya☆'. A yellow warning banner states '利用回数制限あり' (Usage Limit). Below it, text indicates 'Anthropicは提供するAPIを利用します。すべての課金APIを合計して24時間あたり最大100回の利用回数制限があります。[詳細情報]' (Anthropic uses the provided API. All paid APIs are combined with a maximum usage limit of 100 times per 24 hours. [Detailed Information]). The 'Temperature' is set to 0.2, and '参考ページ数' (Reference Page Count) is set to 5. An orange callout bubble points to the usage limit text, containing the text '本日、10回から100回に引き上げ' (Today, increase from 10 times to 100 times).

# おわりに

1. 当初、LLMは日本語に弱いと言われていたが、**古文でもある程度使えるレベル**に改善している
2. 漢字かな交じり文は使えるものの、**かな文字が多い文学作品などは難易度が高い**（自動校訂の必要？）
3. チャットが参照すべき資料を揃える、**「高度な検索システム（レファレンスサービス）」**が重要
4. 生成AIの世界は急速に変化しつつある。持続可能なサービス設計も、**最適解は常に変化する**



日本古典籍 × 生成 AI で過去の文化をもっと身近に

# つくしプロジェクト

Tsukushi project

