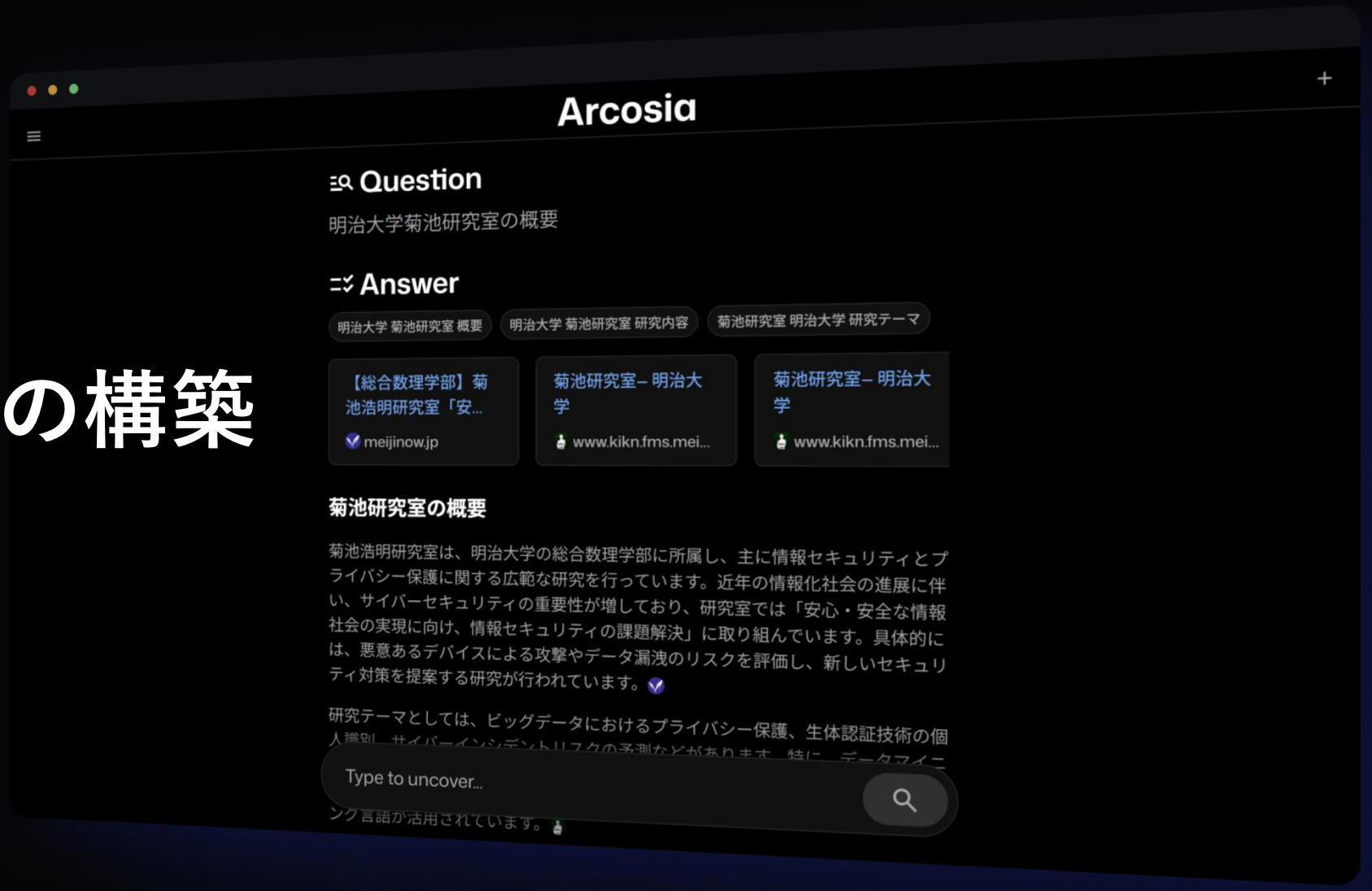
AIを用いた検索エンジンの構築

蓑和　渓太

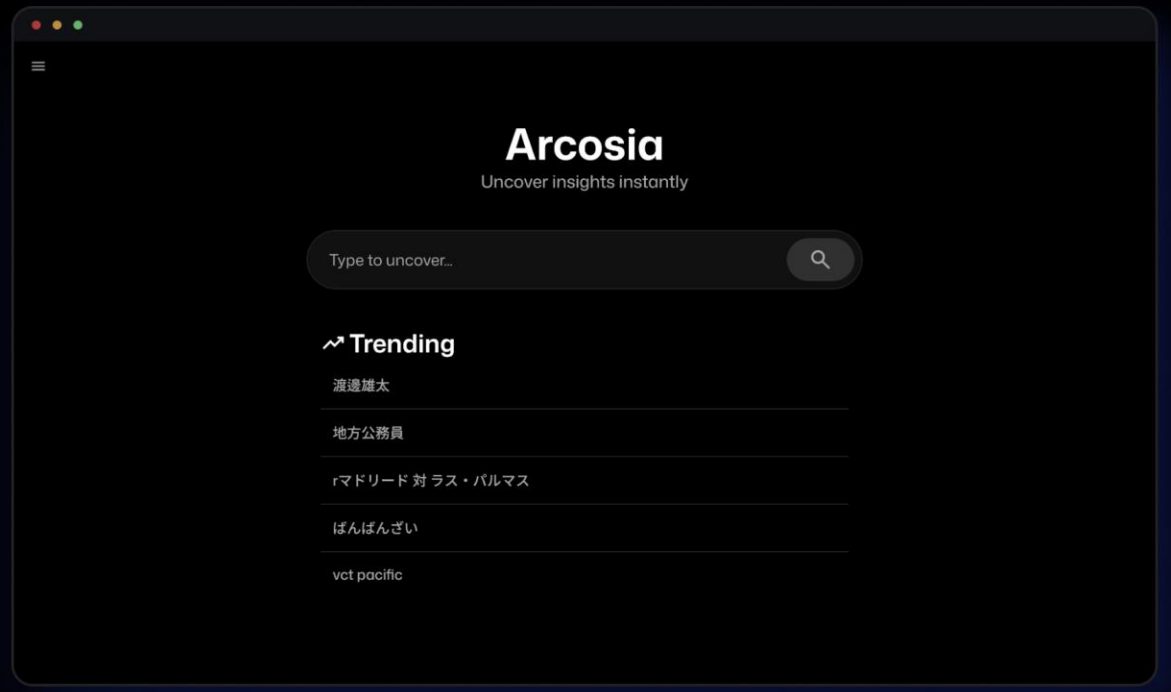


既存のAI検索エンジンの問題点

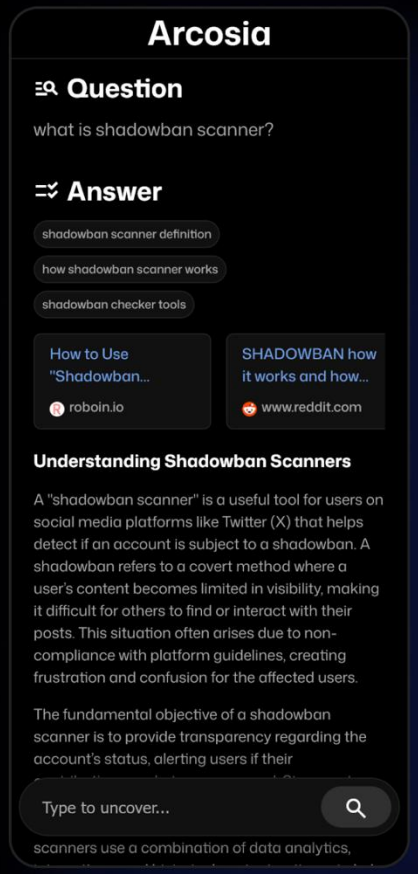
- **費用が高い**
  - ChatGPT Plus：20ドル/月
  - Perplexity Pro：20ドル/月
- **要約精度は高いが検索精度はそれほど高くない**
  - 主要なAI検索エンジンはBingを使っているため
- **PDF非対応？**
  - PDFのアップロードはできるが検索結果のPDFは読み込めない？
- **対象サイトを制限できない**
  - サービスとして展開すると翻案権等の侵害のおそれ

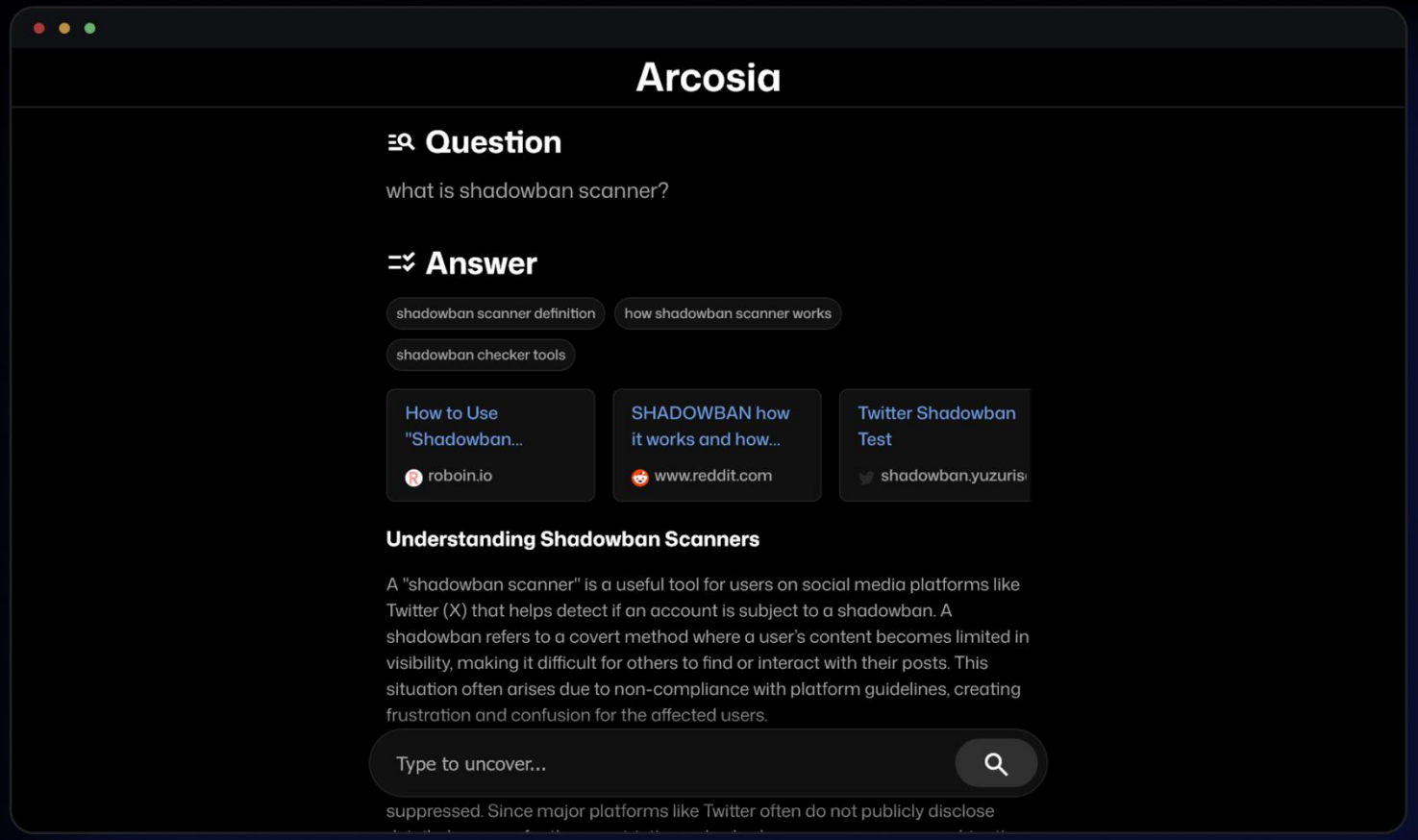
スクリーンショット

Arcosia
Uncover insights instantly


Arcosia
Uncover insights instantly
Type to uncover...
Trending
渡邊雄太
地方公務員
rマドリード 対 ラス・パルマス
ばんばんざい
vct pacific

スクリーンショット


Arcosia
Question
what is shadowban scanner?
Answer
shadowban scanner definition
how shadowban scanner works
shadowban checker tools
How to Use "Shadowban...
roboin.io
SHADOWBAN how it works and how...
www.reddit.com
Understanding Shadowban Scanners
A "shadowban scanner" is a useful tool for users on social media platforms like Twitter (X) that helps detect if an account is subject to a shadowban. A shadowban refers to a covert method where a user's content becomes limited in visibility, making it difficult for others to find or interact with their posts. This situation often arises due to non-compliance with platform guidelines, creating frustration and confusion for the affected users.
The fundamental objective of a shadowban scanner is to provide transparency regarding the account's status, alerting users if their
Type to uncover...
scanners use a combination of data analytics,


Arcosia
Question
what is shadowban scanner?
Answer
shadowban scanner definition
how shadowban scanner works
shadowban checker tools
How to Use "Shadowban...
roboin.io
SHADOWBAN how it works and how...
www.reddit.com
Twitter Shadowban Test
shadowban.yuzuris
Understanding Shadowban Scanners
A "shadowban scanner" is a useful tool for users on social media platforms like Twitter (X) that helps detect if an account is subject to a shadowban. A shadowban refers to a covert method where a user's content becomes limited in visibility, making it difficult for others to find or interact with their posts. This situation often arises due to non-compliance with platform guidelines, creating frustration and confusion for the affected users.
Type to uncover...
suppressed. Since major platforms like Twitter often do not publicly disclose

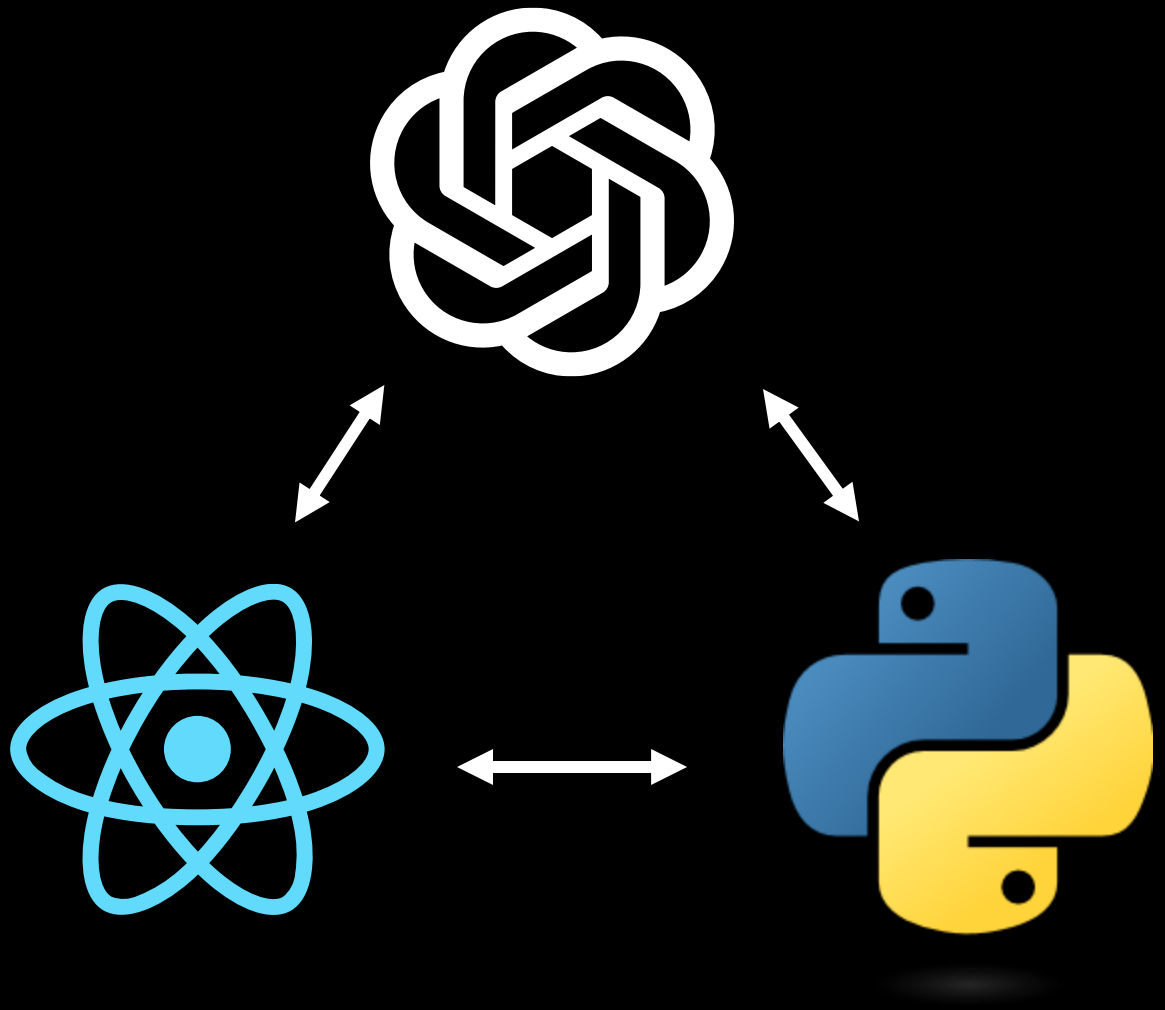
開発したAI検索エンジンの特徴

- **費用が安い**
  - 検索1回あたり1円未満
- **検索精度が比較的高い**
  - Google Custom Search APIを使用
  - 主観評価ではChatGPT search以上Perplexity未満
- **PDFの読み込みに対応**
- **Googleトレンドに対応**
- **必要に応じて自動で検索演算子を活用**
- **サーバーに展開可能（WebベースのUI・HTTPヘッダーによる認証）**

権利保護の取り組み

- robots.txtに対応
- 情報源へのリンクを目立つように表示
- Google Custom Search APIの設定で対象サイトを制限可能
  - →自社サイト内の情報を探索するツールとしての利用など

仕組み



Reactロゴ：https://github.com/facebook/react/blob/main/LICENSE

仕組み

1. ユーザーの質問を分析し、文脈を考慮して3組の検索キーワードを生成
2. Google Custom Search APIで検索
3. 記事をスクレイピング
4. サイドバーなどを除去してメインコンテンツのみを抽出
5. HTMLからMarkdownに変換
6. 文脈を考慮してOpenAI API（GPT-4o mini）で要約
   - 参考サイトのアイコンを表示するためにMarkdown拡張記法を作成。構文木を操作するプラグインを開発してパーサーに挿入
7. サジェストを生成

感想

- **微調整に時間がかかった**
  - コア部分（サーバー側ロジック）の実装：3時間
  - Litを使ったUIの実装：2日
  - Reactへのリプレース：1日
  - 機能の微調整など：3週間弱
- **思ったよりも精度が高い**
- **APIが非常に安い**