

ChatGPT-UserのPDFアクセス観測結果

1. WebPitaの観測結果概要

観測期間：2026-05-05 ~ 2026-05-23

OpenAI系PDFアクセスの約91%がChatGPT-Userによるものだった

本調査期間中におけるOpenAI系クローラーおよびユーザーエージェントによるPDFアクセス合計は **3,725件** でした。内訳は以下の通りです。

項目 (Item)	全体アクセス (Total)	PDFへのアクセス (PDF)	PDF比率 (PDF Ratio)
ChatGPT-User (ChatGPT内取得)	9,067件	3,387件	約37.4%
GPTBot (学習系)	13,125件	181件	約1.4%
OAI-SearchBot (検索用途)	1,729件	157件	約9.1%
OpenAI系PDFアクセス合計	-	3,725件	-

ChatGPT-UserのPDFアクセス比率は約**37.4%**と突出している一方で、GPTBotのPDFアクセス比率は約1.4%、OAI-SearchBotは約9.1%に留まりました。結果として、OpenAI系列によるPDFアクセスの実に約91%がChatGPT-Userによるものであることが判明しました。

2. 観測データに関する考察とサイト特性

今回の観測では、ChatGPT-UserによるPDFへのアクセス比率が非常に高い結果となりました。ただし、この結果には観測対象サイトの特性が影響している可能性を考慮する必要があります。

本データは複数サイトの合算ではなく、単一の法人サイトにおけるアクセスログに基づいています。この対象サイトは、業界動向や国策情報といった専門性の高い情報発信が多く、通常のHTMLページに加え、詳細なPDF資料を多数掲載しているという特徴を持っています。そのため、ユーザーやAIが深い情報を求める行動がこの数値に反映された可能性があります。

3. AIにとってのPDF資料の有用性

なぜChatGPT-UserはこれほどPDFを多く取得しているのでしょうか。その理由として、資料の構造化レベルが挙げられます。

PDF資料は、章立てや見出しが体系化され、報告書、説明資料、仕様書などとして数値、表、定義、参考情報が網羅的に整理されているケースが一般的です。一方、HTMLページはキャッチコピーや概要説明を中心とした、比較的ライトな構成に留まることが少なくありません。AIはより正確で詳細なコンテキストを把握するための「詳細資料」として、PDFを優先的に参照していると推測されます。

4. 仮説：「HTMLで認知し、PDFで詳細を取得する」

今回の結果から、ChatGPTは「HTMLで情報の存在を大まかに把握し、実際の詳細情報はPDFから取得する」という二段階の挙動を行っているという仮説が成り立ちます。

もしこの仮説がGoogleのGeminiやGoogle AI Overviews（生成AI検索）にも共通する特性であるならば、Webサイト側に適切なPDF資料を配置しておくことで、AI Overviewsなどの参照ソースとして選ばれ、表示確率が高まる可能性があります。現時点ではGemini系列のアクセスをChatGPT-Userのように明確に分離して追跡することは困難なため、この仮説の検証には今後も継続的な観測が必要です。

5. 生成AI時代におけるPDF掲載のベストプラクティス

AIに適切に参照・評価されるWebサイトを目指す場合、単にPDFファイルをサーバーに設置するだけでなく、HTMLページとの関係性を明確に設計することが重要です。具体的には以下の施策が示唆されます。

- **正式公開用の適切なPDFのみを掲載する**：構成や情報が古いままのファイルは混乱を招きます。
- **内容を予測できるファイル名にする**：クローラーがコンテキストを理解しやすい名称を付与します。
- **HTMLページ側にPDFの概要を記載する**：検索およびAIが最初に見つける導線を強化します。
- **HTML本文内からPDFへのリンクを明確に設置する**：関連性をクローラーに正しく伝えます。
- **sitemap.xmlに重要PDFを含める**：確実にインデックスを促します。
- **古いPDFや重複するPDFの整理・削除**：評価の分散や古い情報の参照を防ぎます。

これまでPDFは主に対人間向けの「閲覧・印刷用ダウンロード資料」として扱われてきましたが、今後は「AIが参照する高密度な知識ソース」としての役割を併せ持ちます。HTMLとPDFを切り離すのではなく、一体となった情報構造として設計することが求められます。

6. 観測条件および解析対象

本データは、AIアクセス解析ツール「WebPita」が取得したアクセスログから、OpenAIの公開IPレンジ、User-Agent、およびRequest URI情報を基に、以下の条件で抽出・分類したものです。

■ 抽出・分類条件

- **PDF抽出条件**： Request URI に .pdf を含むアクセス。
- **ロボット分類**： GPTBot、OAI-SearchBot、ChatGPT-User、ChatGPT Actions、Visit_From_GPT。

■ 解析対象ログのフィルタリング

ユーザーまたはAIによる実際の「コンテンツ閲覧・取得」行動に絞り込むため、以下のフィルタを適用していません。

- HTTPステータスコードが 200 / 206 / 304 の正常なアクセス。
- 拡張子が html / php / pdf、およびディレクトリURLへのアクセス。
- **除外対象**： 画像、CSS、JavaScript、リダイレクト（301/302）、404エラーなどはノイズとして除外。

7. まとめ

本調査は単一サイトにおける観測データであり、仮説を一般化するにはさらなる複数サイトでの継続的なデータ蓄積が必要です。しかしながら、OpenAI系列のPDFアクセスのうち**約91%がChatGPT-User（ユーザーのプロンプトに応じたリアルタイムな情報取得）に偏っていた**という事実は、今後のWebサイトのSEOおよびAI最適化（AIO）を考える上で極めて示唆に富む結果と言えます。

関連情報・判定詳細：

- WebPitaでのChatGPT関連アクセス判定方法: [webpita_openai_judge.html](#)
- ChatGPTからの送客および送客後遷移の判定について: [chatgpt_web_referral_and_transition_analysis.html](#)

WebPita AI Console for ChatGPT