The Future of Web Data Mining: Insights from Multimodal and Code-based Extraction Methods

Evan Fellman*
Carnegie Mellon Univeristy
efellman@cs.cmu.edu

Jacob Tyo*
DEVCOM Army Research Laboratory
Carnegie Mellon Univeristy
jacob.p.tyo.civ@army.mil

Zachary C. Lipton
Carnegie Mellon Univeristy

Abstract

The extraction of structured data from websites is critical for numerous Artificial Intelligence applications, but modern web design increasingly stores information visually in images rather than in text. This shift calls into question the optimal technique, as language-only models fail without textual cues while new multimodal models like GPT-4 promise image understanding abilities. We conduct the first rigorous comparison between text-based and vision-based models for extracting event metadata harvested from comic convention websites. Surprisingly, our results between GPT-4 Vision and GPT-4 Text uncover a significant accuracy advantage for vision-based methods in an apples-to-apples setting, indicating that vision models may be outpacing language-alone techniques in the task of information extraction from websites. We release our dataset and provide a qualitative analysis to guide further research in multimodal models for web information extraction.

1 Introduction

The extraction of structured information from websites represents a critical challenge in the field of Artificial Intelligence (AI), especially in the context of rapidly evolving web technologies. As the virtual world becomes increasingly central to diverse aspects of society, the ability to efficiently and accurately mine web data is of high importance. This task, commonly known as web scraping, entails navigating the complexities of varied website architectures to extract useful information. The ubiquity of dynamic, visually-rich, and interactive content in modern web design further complicates this landscape, presenting a formidable challenge for automated data extraction technologies.

Historically, web scraping has been dominated by rule-based systems (Gulhane et al., 2011) (Lockard et al., 2018), meticulously designed to accommodate the specific structures of individual websites. The inherent diversity in web design necessitates a tailored approach for each site, significantly limiting the scalability of these systems. Moreover, the dynamic nature of web content, where a single page may present different types of data based on user interaction or other factors such as location or time, adds another layer of complexity. Because of the bespoke nature, rule-based systems often struggle to adapt to dynamic elements, often requiring manual intervention for maintenance and updates.

In the realm of machine learning (ML), the application to web scraping presents unique challenges. The vast differences between websites render the tuning of existing ML systems a daunting task. In most cases, ML-based scraping methods must operate in a zero-shot or few-shot setting, where the model has little to no prior exposure to the specific website from which data is to be extracted. This scenario places a heavy reliance on the innate capabilities of the model to generalize across highly varied environments, a task that has traditionally proven to be challenging for ML systems. As a result, these methods have often been less effective than their rule-based counterparts.

The advent of advanced multimodal AI models has signaled a potential paradigm shift in web scraping methodologies. Pioneering models such as GPT-4 (OpenAI, 2023) and LLaVA (Liu et al., 2023) have demonstrated remarkable capabilities in dealing with complex, multimodal data. These models are equipped to understand and interpret information that spans across text, images, and other web elements, offering a more holistic approach to data extraction. Their prowess in zero-shot performance, where the model can generate useful responses without prior specific training on a task, suggests a significant potential for application in web scraping.

Despite these advancements, the field lacks a

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 1–5

comprehensive and rigorous analysis of such multimodel AI models in the context of extracting practical web data. This gap in research motivates our current study, where we aim to critically evaluate and compare the effectiveness of these cutting-edge techniques in web scraping. Our contributions are as follows:

- A dataset, FanConInfo, of comic convention websites complete with cleaned HTML, a rendered screenshot, and human-annotated labels.
- A rigorous analysis of GPT-4 Vision, GPT-4 Text, and GPT3.5 in extracting information from FanConInfo. We find that leveraging information from a screen capture of a website boosts the accuracy of information extraction by over 20%.
- An error analysis of the methods guiding future work. We find that the vision model predictions align most with human preferences.

2 Related Works

Information extraction from websites has traditionally relied on processing raw HTML code and other text-based structures. Hao et al. (2011) presents a dataset of HTML code with well-defined tasks. For example, on a webpage that describes a book, the dataset asks a system to retrieve the title, author, ISBN-13, publisher, and publish-date using the HTML. Both Hao et al. (2011) and DOM-LM (Deng et al., 2022; Zhou et al., 2021) aim to simplify the DOM tree and feed simplified text embeddings to dense models, achieving state-of-the-art results on benchmarks. More recently, Large Language Models (LLMs) have been used to either directly extract information from website HTML, or to generate a Python program to extract the information from the HTML (Arora et al., 2023). They found this method, (Arora et al., 2023), outpaces methods directly using RoBERTa (Liu et al., 2019) to answer questions, a zero-shot relation extraction method (Lockard et al., 2020) and DOM-LM (Deng et al., 2022). However, these language-only approaches are intrinsically limited when data is stored visually.

Research on pairing vision and language capabilities together in a single model has made rapid progress in interpreting images with text, with models like GPT-4 demonstrating excellent text extraction capabilities from structured documents (OpenAI, 2023), even establishing a new state-of-the-art on the Text Visual Question Answering (TextVQA) dataset (Singh et al., 2019), a dataset designed to challenge model's ability to reason with images. Research is rapid and prolific in multimodal modeling, including the recent work of the multilingual PaLI (Chen et al., 2023) and the modular system of mPLUG-2 (Xu et al., 2023) for multimodal Question Answering (QA).

The dataset by Varlamov et al. (2022) features hand-labeled news articles in raw HTML format, focusing on identifying critical article components like titles and publication dates. Similarly, the Klarna Product Page Dataset (Hotti et al., 2022) contains 51,701 annotated product sale pages for locating key web elements such as buy buttons and prices. Additionally, the Boilerplate Detection using Shallow Text Features dataset (Kohlschütter et al., 2010) includes HTML files labeled to distinguish main content from extraneous elements like advertisements, thus aiding in refining web scraping accuracy. None of the aforementioned datasets provide the ability to compare purely text based and multimodal models on event information extraction.

3 Methodology

3.1 FanConInfo

To enable a fair comparison between visual and textual extraction techniques, we curate a novel dataset, FanConInfo, of comic convention websites which constitute a diverse corpus spanning a range of designs, conventions, and web architectures.

We first extract an initial list of upcoming comic conventions across North America from the aggregator site `FanCons.com`, encompassing fan gatherings to major comic expos. For each convention link, we collect a 3456 x 1878 screen capture and the corresponding HTML content with Selenium (SeleniumHQ, 2023). We remove all CSS styling and `<script>`s from the HTML. Following this, we manually annotate each event with the following attributes: name, start date, end date, and location.

We manually confirmed that when GPT-4 Turbo using the HTML of a webpage and GPT-4 Vision using the screenshot of a webpage agree on the convention name, the name is always correct for the entirety of the dataset. Thus, when the two models agree perfectly, we consider the response as the gold answer. When the models disagree, which oc-

curs 41% of the time across all rows and columns, a human determines the gold response. We only evaluate performance of methods on items that have a label. It is conceivable that some webpages do not list their date nor location, demonstrated in Figure 1, in the above-the-fold portion. In total, our curated dataset contains 86 comic convention websites and is available here.

3.2 Models

For our vision-based model, we leverage the recently released GPT-4 Vision model from OpenAI, `gpt-4-vision-preview` - referred to as GPT-4V. We prompt the model as follows:

> <screen capture placed here>
>
> Get the following information from the given image as a JSON object of strings. Only write the JSON in your response. If any bit is unknown then write N/A instead:
>
> Conference Name: <Name of Conference>,
>
> Start Date: <YYYY-MM-DD>,
>
> End Date: <YYYY-MM-DD>,
>
> Location: <Address or other location>

For our code-based method, we employ the GPT-4 (`gpt-4-1106-preview` - referred to as GPT-4T) and GPT-3.5 (`gpt-3.5-turbo-1106`) models from OpenAI. Rarely when GPT-3.5's sequence length is insufficient to accommodate the entire HTML content, the HTML was truncated. These models were prompted as follows:

> <HTML placed here>
>
> Get the following information from the above HTML as a JSON object of strings. Only write the JSON in your response. If any bit is unknown then write N/A instead:
>
> Conference Name: <Name of Conference>,
>
> Start Date: <YYYY-MM-DD>,
>
> End Date: <YYYY-MM-DD>,
>
> Location: <Address or other location>

3.3 Evaluation

We assess extraction accuracy for 4 key metadata fields: name, start date, end date, and location. We combine the start date and end date into one category. Since the models never deviated from the requested format despite variations on the event pages, a prediction for date is only considered accurate if both are an exact match. To address minor errors, we evaluate predictions for event names and locations using case-insensitive Exact Match (EM) accuracy. Fuzzy matching employs the FuzzyWuzzy Python package (Inc, 2014), measuring:

- Event names: Partial ratio to capture semantic changes with word order (e.g., "ComicCon" vs. "Comic Convention").
- Locations: Partial token sort ratio to allow coherent reordering (e.g., "X Hall, Y Ave., City" vs. "Y Ave., City, X Hall").

This approach balances exact and fuzzy matching for a comprehensive assessment.

GPT	Name	Date	Location	Avg
3.5	0.58(0.05)	0.73(0.06)	0.46(0.06)	0.59
4T	0.62(0.05)	0.74(0.05)	0.56(0.06)	0.64
4V	**0.82 (0.04)**	**0.88 (0.04)**	**0.86 (0.04)**	**0.85**

Table 1: Exact Match accuracy for on the FanConInfo Dataset. The Avg column represents the average accuracy for each model.

	Fuzzy Name		Fuzzy Location	
GPT	Score	Accuracy	Score	Accuracy
3.5	0.88 (0.03)	0.78 (0.05)	0.77 (0.04)	0.62 (0.06)
4T	0.91 (0.02)	0.82 (0.04)	0.83 (0.04)	0.75 (0.06)
4V	**0.95 (0.02)**	**0.92 (0.03)**	**0.95 (0.02)**	**0.94 (0.03)**

Table 2: Partial ratio (name) and partial token sort ratio scores (location) on the FanConInfo Dataset. The score is the average ratio and the accuracy is calculated based on a score threshold of 0.85.

4 Results & Discussion

GPT	Name	Date	Location	Avg
3.5	0.58(0.05)	0.91(0.05)	0.53(0.07)	0.67
4T	0.63(0.05)	**1.00 (0.00)**	0.64(0.07)	0.76
4V	**0.83 (0.04)**	**1.00 (0.00)**	**0.87 (0.05)**	**0.90**

Table 3: Exact Match accuracy for on the FanConInfo Dataset, after removing instances where any of the models predicted that the information is not available. The Avg column represents the average accuracy for each model.

Table 1 shows the visual methodology achieves an average exact match score of 85% while the top text-based methodology achieves an average exact match score of 64%. When relaxing exact match criteria using fuzzy matching, we see the visual methodology achieves an average fuzzy score of 95% when retrieving the convention name while the top code-based method achieves an average fuzzy score of 91% for the same task, as shown in Table 2. When tasked to retrieve the convention

GPT	Fuzzy Name Score	Fuzzy Name Accuracy	Fuzzy Location Score	Fuzzy Location Accuracy
3.5	0.89(0.02)	0.79(0.05)	0.86(0.03)	0.71(0.06)
4T	0.92(0.02)	0.83(0.04)	0.94(0.02)	0.86(0.05)
4V	**0.96 (0.02)**	**0.92 (0.03)**	**0.98 (0.01)**	**0.96 (0.03)**

Table 4: Partial ratio (name) and partial token sort ratio scores (location) on the FanConInfo Dataset, after removing instances where any of the models predicted that the information is not available. The score is the average ratio and the accuracy is calculated based on a score threshold of 0.85.

location, the visual methodology achieves an average fuzzy score of 95% while the top code-based method achieves an average fuzzy score of 83%.

Interestingly, GPT-4 Vision was the highest-performing method across all categories and metrics. Because GPT-4 Vision and Text are the same model, we conclude there exists an advantage when rendering web information as a screen capture in human-readable format versus the traditional HTML machine code.

We also see that it may not always be necessary to use the biggest and most expensive model. GPT-3.5 reaches nearly the same performance as GPT-4 Text, especially when the name is the attribute of interest. This reinforces the advantage of representing web information in human-readable format, as increasing the model capability from GPT-3.5 to GPT-4 had little effect when presenting the model with the HTML representation.

We conducted a comparison between the results of vision-based and code-based methods when both indicate the presence of an answer within the provided mode. The findings are summarized in Tables 3 and 4. Remarkably, even when models express the existence of an answer, the vision-based method consistently delivers more human responses.

4.1 Error Analysis

GPT-4 Vision's errors predominantly come from reading an alternate name prominently displayed, demonstrated in Figure 1. Occasionally interpreting slogans or other emphasized information rather than main headers with event details. However, we do see that the model adapts well to unconventional designs and heavy visual styling, demonstrated in Figure 2. When given only the HTML, the errors tend to primarily originate from missing content, and in some cases, critical information may be exclusively conveyed through images, resulting in issues for models relying solely on the HTML.

Figure 1: Clandestine Comics. GPT-4 Vision read the wrong part as the event name; it predicted "Maryland's Longest-Running Comic Show."

Figure 2: Epic Animation Comic Game Fest. Despite the difficult to read font, GPT-4 Vision was capable of capturing the name. Meanwhile, the date only appears within an image.

Interestingly, we find that when both GPT-4 Text and GPT-4 Vision find a date for an event, Table 3, both methods are correct 100% of the time. The consistent format of dates enables models to achieve high precision in EM.

5 Conclusion and Future Work

In this work, we carry out the first rigorous comparison on practical website data showing strengths of emerging visual approaches versus enduring precision of code for harvesting event details. Our evaluations reveal superior performance in visual-based methods with unparalleled adaptability on designs with heavy imagery. As visual richness accelerates across the web, combining modalities will likely further outpace language-only methods and overcome the shortcomings from unimodal methodologies by blending state-of-the-art coding reasoning with cross-format graphical resilience. Furthermore, we release our dataset to facilitate additional development in event information extraction.

More broadly, our findings posit the understanding of complex webpage images as an important frontier with tangible value for structured data min-

ing from online resources. GPT-4 Vision proves supreme through an average exact match of 85% and fuzzy matching rates of 95% and 95% for name and location data, respectively. We provide strong evidence that rather than competing, effectively integrating textual and visual cues can pave the way for next-generation techniques to achieve new levels of reliability in real-world information extraction across the full diversity of modern web experiences - establishing multimodal web comprehension as a critical area for cross-disciplinary AI development moving forward. Our future work includes expanding this analysis to a wide range of other datasets, including SWDE and the Klarna Product Pages.

References

Simran Arora, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. 2023. Language models enable simple systems for generating structured views of heterogeneous data lakes.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. Pali: A jointly-scaled multilingual language-image model.

Xiang Deng, Prashant Shiralkar, Colin Lockard, Binxuan Huang, and Huan Sun. 2022. Dom-lm: Learning generalizable representations for html documents.

Pankaj Gulhane, Amit Madaan, Rupesh Mehta, Jeyashankher Ramamirtham, Rajeev Rastogi, Sandeep Satpal, Srinivasan H Sengamedu, Ashwin Tengli, and Charu Tiwari. 2011. Web-scale information extraction with vertex. In *2011 IEEE 27th International Conference on Data Engineering*, pages 1209–1220.

Qiang Hao, Rui Cai, Yanwei Pang, and Lei Zhang. 2011. From one tree to a forest: A unified solution for structured web data extraction. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 775–784, New York, NY, USA. Association for Computing Machinery.

Alexandra Hotti, Riccardo Sven Risuleo, Stefan Magureanu, Aref Moradi, and Jens Lagergren. 2022. Graph neural networks for nomination and representation learning of web elements.

SeatGeek Inc. 2014. *fuzzywuzzy: Fuzzy String Matching in Python*.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, page 441–450, New York, NY, USA. Association for Computing Machinery.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Colin Lockard, Xin Luna Dong, Arash Einolghozati, and Prashant Shiralkar. 2018. Ceres: Distantly supervised relation extraction from the semi-structured web.

Colin Lockard, Prashant Shiralkar, Xin Luna Dong, and Hannaneh Hajishirzi. 2020. Zeroshotceres: Zero-shot relation extraction from semi-structured webpages.

OpenAI. 2023. Gpt-4 technical report.

SeleniumHQ. 2023. Selenium. `https://selenium.dev`. Python language bindings for Selenium WebDriver.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Maksim Varlamov, Denis Galanin, Pavel Bedrin, Sergey Duda, Vladimir Lazarev, and Alexander Yatskov. 2022. A dataset for information extraction from news web pages. In *2022 Ivannikov Ispras Open Conference (ISPRAS)*, pages 100–106.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

Yichao Zhou, Ying Sheng, Nguyen Vo, Nick Edmonds, and Sandeep Tata. 2021. Simplified dom trees for transferable attribute extraction from the web.