Optimizing an LLM Prompt for Accurate Data Extraction from Firearm-Related Listings in Dark Web Marketplaces

Conference Paper · December 2024

DOI: 10.1109/BigData62323.2024.10825446

7 authors, including:

Chaido Porlou
Centre for Research and Technology Hellas
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Maria Makrynioti
Centre for Research and Technology Hellas
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Anastasios Alexiadis
Centre for Research and Technology Hellas
42 PUBLICATIONS **323** CITATIONS
SEE PROFILE

Georgios Stavropoulos
Centre for Research and Technology Hellas
47 PUBLICATIONS **301** CITATIONS
SEE PROFILE

Optimizing an LLM Prompt for Accurate Data Extraction from Firearm-Related Listings in Dark Web Marketplaces

Chaido Porlou*‡, Maria Makrynioti*‡, Anastasios Alexiadis*, Georgios Stavropoulos*
George Pantelis†, Konstantinos Votis*, Dimitrios Tzovaras*

*Information Technologies Institute (ITI),
Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
{chaidopor, mmakrynioti, talex, stavrop, kvotis, dimitrios.tzovaras}@iti.gr

†UBITECH LTD, Thessalias 8 and Etolias 10, Chalandri, Athens, Greece, PC 15231
gpantelis@ubitech.eu

***Abstract*—The Dark Web, known for its anonymity and illicit activities, presents considerable challenges for Law Enforcement Agencies (LEAs) due to the complexity and volume of data generated within it. Online marketplaces on the Dark Web are notorious for facilitating illegal activities such as drug trafficking, counterfeit goods, and weapons sales while using advanced obfuscation techniques to avoid detection. The unstructured nature of data on these platforms and their constantly evolving operations make manual extraction and analysis exceedingly difficult.**

This paper addresses the pressing need for structured information extraction from Dark Web marketplaces, with a specific focus on firearm-related listings. Traditional rule-based methods have proven inadequate due to their reliance on HTML tags and pattern recognition, necessitating more adaptive solutions. Thus, the application of Large Language Models (LLMs) and Prompt Engineering to tackle these challenges is explored. By leveraging the capabilities of LLMs, this study aims to transform the extraction process into a more efficient and accurate system. Various generative models and prompt formulations are tested, to determine the most effective approach for extracting detailed information such as product specifications, pricing, and seller details.

The proposed pipeline involves feeding crawled marketplace pages into a generative model, which then identifies Product Details Pages (PDPs) and consequently extracts relevant information from them. The use of LLMs marks a significant advancement over traditional methods, enhancing the accuracy and comprehensiveness of data extraction. Additionally, this research highlights the effectiveness of prompt engineering in improving information retrieval.

This work underscores the critical need for sophisticated tools to monitor and combat illegal activities on the Dark Web, particularly in the context of firearm trafficking. By refining techniques for automated data extraction and applying cutting-edge LLM and prompt engineering methods, this study aims to support LEAs in their efforts to disrupt and dismantle criminal networks and enhance public safety.

***Index Terms*—NLP, NLG, Artificial Intelligence, Prompt Engineering, Large Language Models, Dark Web Marketplaces, Information Extraction, Firearms**

‡Equal contribution.

I. Introduction

The Dark Web has long been a domain of anonymity and illicit activity, posing significant challenges for Law Enforcement Agencies (LEAs) tasked with monitoring and investigating these hidden corners of the internet. As the digital underworld continues to expand, so too does the volume and complexity of the data generated within it. Online marketplaces on the Dark Web are particularly notorious for facilitating a wide range of illegal activities, from drug trafficking to the sale of counterfeit goods and weapons, all while employing sophisticated techniques to evade detection. The unstructured nature of data on these platforms, combined with the constant evolution of their operations, makes it exceedingly difficult for LEAs to extract and analyze relevant information manually. The sheer scale of illegal listings, coupled with the marketplace administrators' deliberate efforts to hinder automated scraping tools, further intensifies these challenges.

Within this context, the extraction of structured information from Dark Web marketplaces has become a paramount objective. The ability to systematically gather and organize data on marketplaces, products, users, and sellers is crucial not only for ongoing investigations but also for developing broader intelligence on criminal networks operating in these spaces. Structured information might enable mapping of relationships between sellers and buyers, identification of trends in illicit goods, and tracing of the movement of these goods across different marketplaces. However, the task is far from straightforward. Traditional rule-based approaches to data extraction, relying on HTML tags and pattern recognition, are often insufficient, requiring extensive manual input and constant adaptation to new data patterns, which is both time-consuming and prone to error. The dynamic and unregulated environment of the Dark Web demands solutions that can cope with the diverse and often deceptive nature of the data found on these platforms.

Among the various categories of illegal products traded on the Dark Web, firearms and firearm-related products represent a particularly grave threat. The trade of illegal firearms online has escalated, allowing criminal organizations and individuals with malicious intent to access weaponry that might otherwise be unattainable through legal means. The nature of firearm listings further complicates matters, as the information is often unstructured, making it challenging to accurately interpret and categorize the content. These listings may include not only complete firearms but also parts, accessories, and ammunition, all of which are critical components of the illegal arms trade. The lack of standardization in how these items are described, along with instances of repeated or incomplete data, further complicates the task of extracting meaningful information.

In light of these challenges, this paper explores the application of LLMs and Prompt Engineering to identify and subsequently extract structured information from firearm-related listings on Dark Web marketplaces. The study leverages the advanced capabilities of LLMs, which have shown remarkable proficiency in understanding and processing online text, to overcome the inherent difficulties posed by unstructured data. The approach involves testing different generative models, to determine the most effective model for this specific task. Additionally, the research explores the domain of Prompt Engineering, systematically evaluating various prompting methods to optimize the LLMs' ability to accurately identify and extract relevant information.

The pipeline developed in this study follows a structured approach. Initially, already crawled and scraped marketplace pages are fed into the generative model. The model is then tasked with determining which of these pages are PDPs of firearm-related products, with the use of carefully fine-tuned prompts. This step is crucial, as it filters out irrelevant or misleading content, focusing the analysis on the most relevant data. Once the desired pages are identified, the model extracts detailed information from each listing, including product specifications, pricing details, and essential seller information such as shipping options and contact methods. This comprehensive extraction process not only captures the explicit details of each listing but might also subsequently reveal underlying patterns and connections that could otherwise go unnoticed.

The use of LLMs represents a significant advancement in overcoming the limitations of traditional data extraction methods. By leveraging advanced Natural Language Processing (NLP) techniques, LLMs can interpret and analyze unstructured data, identifying key details and patterns that would be difficult, if not impossible, to capture using rule-based systems. This approach not only streamlines the process of data extraction but also enhances the accuracy and comprehensiveness of the information gathered, providing LEAs with actionable intelligence to combat the illegal arms trade. Moreover, the flexibility of LLMs allows for continuous adaptation as new data patterns emerge, ensuring that the extraction process remains effective even as Dark Web marketplaces evolve.

This study focuses on developing tools to automate the extraction of structured data from Dark Web marketplaces involved in illegal firearms trade, aiming to help LEAs curb the proliferation of illicit weapons and enhance public safety. Through the testing of various generative models and prompts, the research seeks to provide an effective, automated solution for disrupting criminal networks on the Dark Web, addressing a key challenge in modern digital policing.

In conclusion, as the digital landscape of the Dark Web continues to grow in complexity, the need for sophisticated, adaptable tools for data extraction becomes increasingly urgent. This research not only contributes to the body of knowledge in the field of digital policing but also provides practical solutions that can be implemented by LEAs to enhance their investigative capabilities. By harnessing the power of LLMs and refining the techniques of Prompt Engineering, this study paves the way for more effective and efficient monitoring of Dark Web marketplaces, with a particular focus on curbing the illegal trade of firearms and related products.

II. Related Work

Before exploring the finer details of the research, it's important to first establish a clear understanding of the pre-existing work that has been conducted.

A. Dark Web Marketplaces

Policing the Dark Web poses significant technical, legal, and ethical challenges. Stronger national cybersecurity strategies and international cooperation are crucial to combat illicit activities, as Dark Web investigations face multi-jurisdictional issues and limited data and resources for law enforcement [1]. The Dark Web's dual nature, where both legal and illegal activities coexist, further complicates these efforts. Technologies such as encryption and cryptocurrencies, while protecting privacy and free speech, also facilitate illegal activities, raising complex ethical questions for both law enforcement and researchers [2].

The dynamics of Dark Web marketplaces are shaped by unique mechanisms of trust and reputation. Single vendor shops (SVS) on the Dark Web, for instance, build trust without relying on traditional cryptomarket features like reviews and escrow systems. Instead, these vendors use identity verification, strategic marketing, robust security measures, and other expressions of trust to attract and retain customers. The coexistence of SVS and cryptomarkets suggests that these forms of illicit business have developed sufficient communication and trust signaling mechanisms to thrive simultaneously [3]. Additionally, research has advanced in classifying darknet market listings, particularly those related to firearms. Supervised multi-label text classification techniques have been employed to analyze these listings, demonstrating the superiority of text embeddings and classifiers, such as the TensorFlow Universal Sentence Encoder and SVMs, over simple keyword-based methods [4]. These advancements complement the development of data mining approaches, which enhance the extraction and analysis of patterns within Dark Web data, offering valuable insights for law enforcement [5].

The development of methods for downloading and analyzing Dark Web content has seen significant progress, particularly in creating reliable approaches that address the challenges of capturing consistent data. These methods have improved the ability to study Dark Web marketplaces in detail [6]. Complementary to this, an analytical framework utilizing free tools like AppleScript has been introduced to automate Dark Web scraping and analysis, providing a more accessible alternative to traditional web scraping techniques, albeit with some limitations, such as the need for frequent script adjustments due to changes in data structure [7]. Furthermore, the public release of a dataset on cybersecurity-related listings from Dark Web marketplaces has provided researchers with valuable resources. Collected over several years, this dataset reveals the accessibility and pricing of illegal cyber products, making it a crucial tool for understanding the dynamics of these markets without direct interaction [8].

Research into the pricing and availability of illicit firearms on the Dark Web has provided critical insights into an underground market. An analysis of firearm vendors revealed that long guns are generally priced below their MSRP, while handguns are often priced higher. Vendor characteristics, however, play a relatively minor role in pricing [9].

B. Firearm Trafficking on Marketplaces

Further research has expanded the understanding of illicit arms trade on the Dark Web through crime script analysis, which details the step-by-step processes vendors use to advertise, sell, and deliver illicit firearms. This approach offers a comprehensive view of the roles and functions of participants within a market [10]. The concentration of weapons trafficking within a few major cryptomarkets has also been highlighted, with innovative research cross-referencing pseudonyms with online traces, such as PGP keys, to provide more accurate estimates of vendors involved in weapons sales [11]. In addition, ensemble machine learning models have been employed to detect illegal firearms listings, significantly aiding law enforcement in identifying key vendors within the illegal firearms ecosystem. This method enhances the classification of firearms-related content, providing actionable intelligence for law enforcement interventions [12].

Pioneering efforts in counter-terrorism have also been made by applying machine learning classification techniques to detect modern weapons procurement by terrorist groups on Dark Web forums. This approach, utilizing expert-annotated posts and social media analytics, offers the first automated model for identifying weapon procurement activities in online environments, marking a significant advancement in counter-terrorism strategies [13]. A qualitative analysis of illicit arms trafficking has also been conducted using a crawler based on the ACHE Python library, which collected data from ten Tor network marketplaces. This study provided a comprehensive overview of the illegal arms trade, including the types of weapons available, payment and shipping methods, and vendor practices, aligning with findings from institutions like the United Nations and RAND Europe [14].

Finally, research into the availability of weapons across 20 darknet markets has provided a snapshot of the scale of illicit arms trafficking online. Between July and December 2019, over 2,100 weapons were identified, highlighting the extensive range of items available for sale and the presence of vendors operating across multiple markets [15] .

C. Marketplace Information Extraction

The application of web scraping and information extraction techniques in the context of Dark Web marketplaces has seen significant advancements. A methodological framework has been introduced to ensure the validity of web data collected through scraping and APIs. By addressing both technical and ethical challenges, this framework provides a typology of web data use in marketing research, offering new directions for capturing evolving marketplace realities [16]. In parallel, a web scraping application developed using Python and BeautifulSoup has been created to identify and report illegal sales of branded products on e-commerce platforms. This application achieves notable precision and recall rates, demonstrating its effectiveness in detecting unauthorized sales [17].

Furthermore, a comprehensive overview of web scraping techniques has highlighted their critical role in giving E-Commerce companies a competitive edge, emphasizing the strategic benefits of these methods for market analysis, price optimization, and customer sentiment analysis, while also navigating the legal and ethical considerations inherent in such practices [18]. Moreover, a study focusing on the Data Marketplaces (DMs) ecosystem has revealed unknown details about the pricing and features of data products listed across various platforms. This study underscores the challenges in comparing prices and developing a quotation tool for data products, revealing significant variations in the data market landscape [19].

To enhance entity extraction accuracy, a Named Entity Recognition (NER) model has been introduced, leveraging a Bi-directional LSTM with CNN architecture. This model, specifically targeting product information on the Dark Web, has set new benchmarks in accuracy, outperforming existing methods and addressing the limitations of previous NER solutions in the e-commerce domain [20]. The application of LLMs to various tasks, including HTML understanding and information extraction, has demonstrated their exceptional performance. By fine-tuning LLMs on HTML-specific data, a study showed significant improvements in classification accuracy and task completion rates, advancing the capabilities of LLMs in web automation and retrieval tasks [21].

D. Prompt Engineering with LLMs

Prompt engineering has also emerged as a critical area of research, with a comprehensive survey highlighting 39 different prompting methods used across 29 NLP tasks. This research underscores the transformative impact of prompt engineering on LLM performance, offering a detailed taxonomy that sets the stage for potential state-of-the-art approaches in NLP [22]. Further exploration into prompt engineering has revealed

its role in extending the capabilities of LLMs and vision-language models (VLMs) without modifying core parameters. By categorizing and analyzing various prompting methods, this research addresses a critical gap in understanding this rapidly evolving field, offering a valuable resource for future research [23] .

In the context of information extraction, a locally deployed LLM with prompt engineering has been utilized to transform these tasks into dialog-based processes. This approach has achieved high accuracy in extracting material entities and types, surpassing fine-tuned methods despite certain limitations in recognizing long entities [24]. Similarly, another study has focused on extracting structured information from unstructured text using a prompt engineering approach with conversational LLMs. This method simplifies information retrieval and analysis, offering a groundbreaking approach to managing unstructured text in a way that makes it significantly easier to query and analyze [25] .

Additionally, the effectiveness of different prompt engineering methods for knowledge extraction has been explored, with findings indicating that incorporating a single relevant example into prompts can significantly enhance performance. This study introduced a novel evaluation framework based on Wikidata ontology, revealing that retrieval-augmented prompts are more effective than reasoning-oriented methods in certain contexts [26] . Lastly, an application leveraging Generative AI (GenAI) and LLMs within a retrieval-augmented generation (RAG) architecture has been presented to automate information retrieval from textual data in PDFs. This application significantly reduces the time spent on manual research tasks, offering a scalable and efficient solution for document analysis [27].

This paper introduces an approach to Dark Web monitoring by harnessing the power of LLMs and prompt engineering, offering an adaptive alternative to traditional static, rule-based scraping methods. Unlike previous models reliant on rigid data parsing, this framework uses LLMs to extract detailed, contextually relevant information from unstructured marketplace data, effectively identifying and classifying firearm listings while capturing details like product specifications and pricing. By implementing tailored prompts that adapt to the nuances of firearm-related data and marketplace changes, this system is capable of handling evolving terminology without frequent recoding. Additionally, the integration of LLMs to automatically analyze PDPs enhances the accuracy of data capture, enabling law enforcement agencies to obtain actionable intelligence at scale. This approach represents a significant advancement in the automated monitoring of illicit online activities.

III. Data Collection & Preprocessing

This section provides a detailed overview of the data collection process and the preprocessing steps applied to the data.

A. Data Collection

In the initial stage of this study, the emphasis resides on data collection using pre-established crawling and scraping techniques. Data collection is critical, as the quality and comprehensiveness of the data directly influence the accuracy of the findings. The challenge of collecting data from Dark Web marketplaces, which are often deliberately concealed or inaccessible, requires a careful and methodical approach.

An assortment of online Dark Web marketplaces has been chosen, crawled, and scraped prior to the experiments of this study, by the company and CEASEFIRE partner Ubitech [28]. The quantity of marketplaces is equal to 6, and their characteristics can be seen in Table I. This section details the sources of the aforementioned data, the process of its collection, and the characteristics of the dataset used in this analysis.

TABLE I
MARKETPLACE DATA STATISTICS

Marketplace	Page Number	Mean Page Word Count	Size
DarkDeepMarketplace	5000	746.29	525 kB
Digital Thrift Shop	1405	297.40	104 kB
Venus Marketplace	501	643.60	29 kB
Drkseid	18	75.33	92 kB
BlackMarketGuns	501	569.76	109 kB
TorGuns	56	267.71	552 kB

The process of data collection involved manually searching for and identifying Dark Web marketplaces that feature firearm-related listings. The search began by exploring known directories and forums frequented by Dark Web users, which often serve as hubs for discovering active marketplaces. Various keywords related to firearms, weaponry, and related products were employed to locate relevant marketplaces, using Dark Web search engines like Ahmia. Marketplaces were selected based on the presence of firearm-related listings, since most of them prohibit the upload of such content. Once identified, the marketplace URLs were cataloged for further analysis and data extraction.

As mentioned before, the primary dataset for this study was obtained from Ubitech, a provider specializing in online data collection. The dataset includes records from 6 distinct marketplaces, encompassing a wide array of illegal goods and services. This data was collected in October 2023, resulting in a comprehensive snapshot of marketplace activity on that month. The data is stored in a structured JSON format, which facilitates subsequent analysis.

To gather the necessary data, a sophisticated web scraping tool was deployed to systematically navigate and extract content from targeted Dark Web marketplaces. The tool accesses marketplaces through the Tor network, ensuring anonymity and secure data retrieval. Scraping is subsequently performed in order to include Product Listing Pages (PLPs), Product Detail Pages (PDPs), user and seller profiles, and pages with general marketplace information. The data is then exported in a standardized format, ensuring consistency across different marketplaces.

The format of a scraped marketplace page can be seen below:

```
{   "url": URL,
    "html_raw": RAW HTML WEBPAGE,
    "timestamp": TIMESTAMP  }
```

B. Data Preprocessing

Before any analysis can be performed, the raw data collected from Dark Web marketplaces requires thorough preprocessing. This step is essential to remove irrelevant content, clean the data from any potential noise, and ensure that it is in a format suitable for natural language processing tasks. The preprocessing stage lays the foundation for accurate and efficient information extraction in the subsequent phases of the study.

The raw HTML data extracted from Dark Web marketplaces contains a number of tags, scripts, and other non-essential elements that can interfere with the performance of language models. To address this, we utilize BeautifulSoup [29], a Python library, to parse the HTML and remove all redundant tags. This cleaning process not only reduces the data's complexity but also standardizes the text, making it more suitable for processing by the language models employed later.

The removal of HTML tags from the collected data is a crucial step in the preprocessing pipeline. Raw webpages often contain a variety of HTML tags that are irrelevant for language model analysis. These tags can interfere with the model's ability to process and understand the text by introducing elements that are not part of the actual content. By stripping out these tags, the data is rendered in a clean, text-only format that aligns with the requirements of LLMs, the main tool utilized in this study. This transformation ensures that the LLM can focus on analyzing and extracting meaningful information without the distraction of non-textual elements.

Furthermore, cleaning HTML tags significantly contributes to noise reduction. Removing tags leads to a more refined dataset where only the core textual content remains. This refinement is especially important for subsequent processing stages, such as tokenization and use of word embeddings. In a standardized text format, the LLMs can more effectively tokenize the text—breaking it down into meaningful units—and generate precise word embeddings; numerical representations that capture the semantic meaning of each word. The result of this process is a dataset where words are isolated from HTML structures, making it more suitable for analysis and enhancing the accuracy of information extraction by a generative model.

IV. Methodology

This section outlines the methodology used to conduct the research and establish the experimental setup of the study.

A. Prompt Engineering for Data Filtering

Once the data has been curated for use as prompting input, the next step is to identify and filter firearm-related product pages that contain individual product listings. In e-commerce, these single-product pages are typically referred to as Product Detail Pages (PDPs), where individual items are presented with detailed information, such as descriptions, specifications, pricing, and other relevant details. This distinction is critical because the information required for analysis is most comprehensively available on these single-listing pages, as opposed to category or multi-product pages.

The task of identifying single listings, however, presents several challenges. One of the main difficulties arises from the ambiguity in the structure of marketplace pages. Many pages may appear similar when they showcase single products or multiple listings, making it hard to differentiate between a PDP and a Product Listing Page (PLP). Moreover, marketplace pages are frequently dynamic, with content that can change or refresh based on user interaction or system updates. Such changes can disrupt static identification techniques. Additionally, overlapping content, such as the display of recommended or related items on a single product page, can further blur the line between a true single listing and a multi-listing format. These issues make it challenging to reliably identify PDPs based on page structure alone.

In order to tackle this first part of the pipeline, a prompt has been meticulously crafted and evaluated. More specifically, the prompt requests the categorization of the input page as a PDP or a non-PDP, hence it is a classification prompt. The input for an LLM combines a question, which directs the model's response, and a system prompt, which provides the cleaned HTML. Together, these elements guide the model in generating relevant and coherent answers. The final prompt question can be seen below:

```
"Answer to this question with yes or no: Does
this url and cleaned html belong to a marketplace
page focusing on one single product listing about
a firearm or a firearm-related product? ATTENTION,
if the page directs to add to cart, add to wishlist
pages or tag pages then answer with NO."
```

The models that have been tested for this specific sub-task are the: llama3-70b-8192, llama3-8b-8192 [30] and gemma-7b-it [31]. Meta released the Llama 3 family of LLMs in 8B and 70B sizes, optimized for dialogue and outperforming many open-source chat models on industry benchmarks. Google's Gemma models, built from the same research as the Gemini models, are decoder-only LLMs available in English, designed for tasks like question answering and summarization. Both Llama 3 and Gemma offer pretrained and instruction-tuned variants, focusing on adaptability and high performance across text generation tasks.

B. Prompt Engineering for Information Extraction

The process for designing and refining prompts to extract structured information from firearm-related listings on Dark Web marketplaces involves careful prompt engineering and testing with various LLMs, more specifically the models llama3-70b-8192, llama3-8b-8192 and gemma-7b-it, as mentioned before. These models are employed to ensure effective extraction across different formats and complexities of product listings. Furthermore, the prompts are tailored to handle Product Detail Pages (PDPs), ensuring accurate extraction of relevant entities while maintaining flexibility across various page formats.

The design of the prompts is based on the desired output structure, which is in a key-value format. Each listing's extracted details are organized into the predefined entities shown below. This structure serves as a template for all extraction tasks and ensures uniformity in the outputs generated from the LLMs.

```
{   "listing_title": LISTING TITLE,
    "product_brand": PRODUCT BRAND,
    "product_type":  PRODUCT TYPE (PISTOL ETC.),
    "product_specs": PRODUCT SPECIFICATIONS,
    "product_price": PRODUCT PRICE,
    "product_quantity": AVAILABLE PRODUCT QUANTITY,
    "shipping_info": SHIPPING INFO,
    "contact_info": SELLER CONTACT INFORMATION    }
```

The prompts are categorized based on the prompting methods. The prompt categories that are studied are the: Standard Prompt, Expertise Prompt and Itemized Prompt. The Standard Prompt includes a request of no extra sentences before and after the structured response, a mention of the desired dictionary format and the individual product information to be extracted, and the mention of a None addition if an entity does not exist. The Expertise Prompt builds on the Standard Prompt by adding a layer of expertise, portraying the model as both a Law Enforcement Agent (LEA) and a Named Entity Recognition (NER) specialist, giving the prompt a more authoritative tone. Finally, the Itemized Prompt builds on the Standard Prompt by incorporating specific rules for each entity, aiming for a more detailed and accurate extraction. These prompting methods have been selected to ensure a structured and efficient approach to the information extraction task execution, with further details on each provided below.

Standard Prompt: A Standard Prompt is a concise instruction that provides just enough detail to guide the information extraction process without overwhelming the system with extra information. This method emphasizes clarity and simplicity, focusing on extracting information in a structured way, such as using a dictionary as the output format. It avoids unnecessary text and assumes no missing values, making the process more efficient and less ambiguous.

Expertise Prompt: An Expertise Prompt involves assigning specific roles to the model to enhance its performance in a given context. For instance, in this approach, the model could be directed to assume the roles of both a Natural Language Processing (NLP) Expert and a Law Enforcement Agent (LEA). As an NLP Expert, the model would focus on understanding and applying advanced technical methods for data processing and analysis. In the role of an LEA, the model would emphasize accuracy and attention to detail, considering the practical implications of the task. By integrating these roles, the model's performance is tailored to address both the technical intricacies and the critical importance of detail within the context of the task.

Itemized Prompt: An Itemized Prompt provides a detailed breakdown of the specific entities to be extracted, along with their expected results and examples. This method involves a comprehensive explanation of each component required for the task, ensuring that all relevant aspects are addressed. For example, an Itemized Prompt might specify that the model should identify and extract entities such as title, price, and specifications from a page. It would include clear descriptions and examples of how these entities should be recognized and processed. This detailed approach ensures that the model understands exactly what is required, facilitating more accurate and precise outputs.

An LLM's input consists of two parts: a system prompt that delivers the cleaned HTML and a question that guides the model's response. These components work together to direct the model in producing valuable and correct extraction. Each method and its respective prompt question is presented in Table II.

C. Post-processing Method

In addition to trimming unwanted content, the post-processing method also includes a filtering step to ensure that PLPs are excluded. The LLM is designed to filter this information out, but in some cases, Product Listing Pages containing multiple entries may bypass initial preprocessing filters (section IV-A). To detect these cases, a regex search within the output to identify instances where more than one structured listing has been extracted is performed. If multiple listings are found, the entire page is flagged and filtered out. This ensures that only PDPs are retained for further analysis, improving the consistency and accuracy of the data extraction process.

In the post-processing stage, the focus is on ensuring that the output generated by the LLMs adheres to the predefined structured key-value format. Although the LLM is provided with clear instructions to generate only the structured data that was analyzed in the previous subsection for each listing, there are instances where it fails to fully comply with these instructions. The model sometimes includes unstructured content, either before or after the key-value output, often adding unnecessary explanations or extra text. To address this issue, a trimming process is employed using regular expressions (regex). This technique isolates and retains only the relevant structured data, removing any extra content to maintain the consistency and accuracy of the extracted information.

By combining filtering out PLPs and regex-based trimming to remove unstructured content, the post-processing phase acts as a crucial layer of quality control. This ensures that the final dataset consists solely of well-structured, single-product listings, ready for further analysis.

V. Results

In this section, the evaluation focuses on two key aspects of the performance of various LLMs and prompt configurations. First, the ability of different LLMs to accurately filter PDPs from Dark Web marketplaces is assessed, using metrics such as Accuracy, Precision, Recall and F1-score. Secondly, the evaluation examines the effectiveness of different models and prompting methods in extracting structured information

TABLE II
PROMPTING METHODS

Standard Prompt Question	Expertize Prompt Question	Itemized Prompt Question
"I want information to be extracted in a valid dictionary format from a cleared from tags HTML text. The entities are `listing_title`, `product_brand`, `product_type`, `product_specs`, `product_price`, `product_quantity`, `shipping_info`, *and* `contact_info`. *Please don't stray from the provided text in your answers. Please don't answer with unnecessary text, only the dictionary in a valid form. Focus on the main product of each page. If you can't find a suitable answer, fill the value with None. The specifications value must be the only one in nested dictionary form."*	*"You are a Named Entity Recognition expert tasked with extracting specific entities from HTML pages for a Law Enforcement Agency task. You need to extract the following entities from each page and return them in a valid structured dictionary format:* `listing_title`, `product_brand`, `product_type`, `product_specs`, `product_price`, `product_quantity`, `shipping_info`, *and* `contact_info`. *- Focus primarily on identifying the main product per page. - The value of the* `product_specs` *must be in a valid dictionary format. - Extract only the dictionary in a valid format; do not infer or add extra commentary. - Strip all HTML tags and work only with the cleaned text data. - If no entities of a given type are found, return 'None' for that type. - Avoid ambiguity, partial information, or incomplete entities unless explicitly stated in the text."*	*"I want information to be extracted in a valid dictionary format from a cleared from tags HTML text. Please don't stray from the provided text in your answers. Please don't answer with unnecessary text, only the dictionary in a valid form. Focus on the main product of each page. If you can't find a suitable answer, fill the value with None. The specifications value must be the only one in nested dictionary form. The entities are: -* `listing_title`: *Find the unique title that appears on the listing; this should be one unique product. -* `product_brand`: *Find the brand; it can be the manufacturer that commonly appears on the title. -* `product_type`: *Find the product type of the product, for example, rifle, pistol, etc. -* `product_specs`: *Find all the specifications of the product; they commonly appear in a list, but if you find more in the text, add them to the list. Create only one dictionary for the value; don't nest any. -* `product_price`: *Find the price after the sale, and keep the dollar sign. -* `product_quantity`: *Find how many pieces of the products are available. -* `shipping_info`: *Find information about the shipping, for example, geographical info, duration of shipping, delivery charges. -* `contact_info`: *Find possible contact details, for example, social media handles, emails, phone numbers, etc."*

from these pages, using Exact Match and Similarity Score as performance metrics.

For the purpose of the above evaluation, a dataset was created consisting of 30 carefully selected pages from the Dark Web marketplaces discussed in section III-A. This collection includes a variety of pages, such as Product Detail Pages, Product Listing Pages, and other irrelevant marketplace pages, to ensure a comprehensive evaluation. Each instance in the dataset consists of a URL, a clean version of the page text with no HTML tags, and a label indicating whether it is a PDP or non-PDP. For each PDP, the dataset also specifies the information that should be extracted, as detailed in section IV-B. This dataset was manually curated and annotated to provide a reliable basis for assessing the performance of different LLMs and prompt methods in both filtering and information extraction tasks.

A. PDPs Filtering Evaluation

In the first evaluation stage, the focus is on evaluating the ability of different LLMs to filter out PLPs and other irrelevant marketplace pages, and retain only PDPs. The accuracy of this filtering process is critical, as any retained PLPs or irrelevant pages can distort the results during the extraction process. The performance of the models is measured through accuracy, which reflects how well each model identifies and isolates PDPs, and response time, which captures the computational efficiency of the filtering process.

Table III provides the accuracy, recall, precision and f1-score metrics for each model. The results indicate that the model llama3-70b-8192 exhibits the highest accuracy score of 96.7% on the evaluation dataset. It is noted that the model llama3-8b-8192 is a close second in both f1-score and accuracy, while it has a perfect score of recall. On the other hand, the model gemma-7b-it, while having a relatively high precision, falls behind on the recall metric. This translates as the model's inability to consistently categorize single listing pages as PDPs, and instead flags them as PLPs or irrelevant.

TABLE III
FILTERING EVALUATION

Model	Precision	Recall	F1-Score	Accuracy
llama3-70b-8192	**0.952**	**1.000**	**0.976**	**0.967**
llama3-8b-8192	0.870	**1.000**	0.930	0.900
gemma-7b-it	0.917	0.550	0.687	0.667

B. Info Extraction Evaluation

The Info Extraction task is evaluated using a combination of Exact Match and Similarity Score to assess the models'

overall performance. Exact Match measures the percentage of extracted data that perfectly aligns with the ground truth, while Similarity Score evaluates the degree of closeness between the extracted data and the expected output, offering a more flexible understanding of model accuracy. More specifically, Similarity Score expresses the cosine similarity between the ground truth and the extracted data, by using embeddings from the DarkBert model [32]. In both metrics, a mean percentage is calculated and used for the final evaluation.

It is important to note that the *product_specs* entity poses a relatively high challenge for evaluation. Specifically, the manual annotation of the specifications for each PDP often differs significantly from the model's extracted specifications, due to factors like nested dictionaries or varying levels of detail. As a result, the Exact Match metric always returns 0 for the specifications, as it fails to account for even minor differences. To address this, a combined metric has been developed: all entities are evaluated using the Exact Match metric, except for specifications, which are assessed using the aforementioned Similarity Score. The average of the 8 entities is then calculated for each PDP, following the same approach as the individual metric calculation. Table IV presents the results for each model and prompting method combination, detailing performance across these three key metrics.

From a strict evaluation perspective, the results indicate that the Standard Prompt consistently achieved an accuracy of 34-38% when assessed using the Exact Match method throughout the models. Notably, the Expertise Prompt question yielded the highest performance, reaching an accuracy of 39% with llama3-70b.

From a contextual perspective, the similarity evaluation revealed that all prompts demonstrated similar accuracy, with each achieving an impressive 81%. The Itemized Prompt exhibited slightly lower performance at 79% when evaluated using the llama3-70b model.

The gemma-7b-it model, when using the Expertise Prompt question, was unable to generate a result, responding with: *"The provided text does not include any HTML pages or product information, so I am unable to extract the requested entities. Therefore, I cannot provide the requested data."*

Overall, the combination of Standard Prompt with the llama3-70b not only achieves the highest Exact Match of 38% but also records a high Similarity Score of 81%, indicating strong alignment with the ground truth across a range of listings.

VI. Discussion

In this section, the key challenges, insights, and implications of the research will be discussed in depth. This includes an analysis of the effectiveness of LLMs for structured information extraction, reflections on the limitations encountered, and considerations for future improvements and research directions.

A. Analysis of Findings

In this study, the effectiveness of LLMs for extracting structured information from Dark Web marketplaces is thoroughly examined. These models, designed for versatile natural language processing tasks, exhibit varying performance levels in the marketplace information extraction task.

Among the models evaluated, the best performers are identified using three key metrics: the Exact Match evaluation metric, a Similarity Score based on word embeddings from DarkBert, and the Specs-Adjusted Exact Match metric. In terms of the Exact Match metric, model llama3-70b-8192 demonstrates the highest performance, achieving an accuracy score of 39%. This score is fairly adequate for a metric as strict as the Exact Match, where even minor details—such as articles, measurement units, or additional specifics—can cause an otherwise correct output to be deemed incorrect. In contrast, it is evident that the Similarity Score for almost all models is relatively high, at 81%. This metric is more suitable for the task at hand, as it emphasizes the holistic nature of information extraction. If the extracted information closely aligns with what is presented on the page, the task can be considered significantly successful. It is important to note that care should be taken to ensure the model focuses solely on information extraction and does not generate or input information on its own. This aspect can sometimes be overlooked during evaluation with the Similarity Score metric.

Last but not least, the highest score on the Specs-Adjusted Exact Match metric is also achieved by the llama3-70b-8192 model, at 52%. As mentioned before, the use of this metric highlights the challenges associated with evaluating the *product_specs* entity. The 52% score on the Specs-Adjusted Exact Match metric, compared to the 39% score on the standard Exact Match, emphasized the importance of this combined evaluation approach. This improvement indicates that the Similarity Score offers a more realistic measure of performance, especially given the complexities of nested structures and varying detail levels. The increase in score suggests that capturing the essence of specifications may be more valuable than strict adherence to exact outputs. This disparity emphasizes the need for flexible evaluation metrics in areas where strict matching is challenging.

Furthermore, it becomes evident that prompt engineering also plays a crucial role in optimizing model performance. The results show that the Standard Prompt method consistently outperforms others across all models in the Exact Match metric, achieving scores of 38%, 34% and 34%. This consistency emphasizes the effectiveness of clear, concise and fairly simple instructions in improving strict adherence to expected outputs. Notably, all prompting methods yield the same Similarity Score, as mentioned before. The Expertise Prompt, which combines the roles of an LEA and an NLP expert, achieves a notable 52% on the Specs-Adjusted Exact Match metric. The LEA perspective prioritizes compliance and accuracy, while the NLP expertise improves the model's ability to identify nuanced distinctions in specifications. By guiding the model to focus on critical attributes such as dimensions, materials, and features, this tailored prompt enhances extraction accuracy. This demonstrates the importance of integrating specialized roles in prompt design to boost performance in complex

TABLE IV
INFO EXTRACTION EVALUATION

LLM	Prompting Method	Exact Match	Similarity Score	Specs-Adjusted Exact Match
llama3-70b-8192	Standard Prompt	0.38	**0.81**	0.51
llama3-70b-8192	Expertise Prompt	**0.39**	**0.81**	**0.52**
llama3-70b-8192	Itemized Prompt	0.16	0.79	0.28
llama3-8b-8192	Standard Prompt	0.34	**0.81**	0.47
llama3-8b-8192	Expertise Prompt	0.33	**0.81**	0.45
llama3-8b-8192	Itemized Prompt	0.24	**0.81**	0.37
gemma-7b-it	Standard Prompt	0.34	**0.81**	0.47
gemma-7b-it	Expertise Prompt	-	-	-
gemma-7b-it	Itemized Prompt	0.32	**0.81**	0.44

information extraction tasks. Additionally, it is worth noting that the gemma-7b-it model's inability to generate a result when using the Expertise Prompt is likely due to ethical concerns and the model's inherent reluctance to engage with sensitive or illegal content.

Another notable observation is that the traditional Chain of Thought prompting method, which generally involves breaking down a task into smaller, sequential steps, is not ideal for processing a high volume of pages efficiently. In comparison, the proposed prompting methods treat each HTML page as an input to the LLM, enabling the extraction of all relevant data in one pass. Furthermore, it is important to mention that LLMs are constrained by their maximum input character limits, which poses a challenge when processing extensive HTML content. Given that each page of cleaned HTML text can be quite lengthy, the traditional few-shot technique—which involves providing incremental examples and building context progressively—was deemed impractical. Instead, the experiment focuses on directly inputting entire HTML pages into the LLM along with specific prompt directions. By doing so, the extraction process is streamlined and performance is enhanced, especially when dealing with substantial amounts of marketplace content.

Finally, post-processing also emerges as a critical phase in this pipeline. Despite detailed instructions, LLMs often produce outputs that are only partially structured, failing to follow the predefined guidelines to the letter. This inconsistency suggests that while LLMs can be highly effective, they sometimes require post-processing to ensure the extracted information aligns with the expected structure. This phase involves cleaning and restructuring the output to make it usable for subsequent analysis. Therefore, post-processing should not be overlooked as it significantly enhances the quality of the results by compensating for the models' tendency to overlook certain instructions.

B. Challenges

Several significant challenges emerge during this study. A key issue observed was that the models often "refused" to extract certain types of information, particularly when dealing with sensitive or illegal content, such as listings related to weapons. This behavior stems from the models being trained on datasets that incorporate ethical guidelines or restrictions, which discourage engagement with illicit material. As a result, some information may have been misinterpreted or not extracted at all, introducing gaps in the final dataset. This legal and ethical filtering, while well-intentioned, presents a challenge in gathering comprehensive data from these marketplaces and complicates the models' performance in contexts where complete information extraction is crucial for analysis.

Another key issue observed was the absence of a well-suited evaluation metric for measuring the effectiveness of the LLMs in this specific task. Standard LLM evaluation metrics (BLEU, ROUGE etc.), while useful, are not fully capable of capturing the requirements of extracting structured information from unstructured marketplace data. As a result, there is some difficulty in accurately assessing the models' performance and their ability to meet the specific needs of this study. Additionally, as already mentioned, the LLMs sometimes produce results that deviate from the desired format, particularly when dealing with the complex structures of some marketplaces. Each marketplace presents unique challenges, and the inability of the models to adapt consistently across different, mainly specification-related, formats is a notable limitation.

C. Future Work

In the realm of future work, there are several promising avenues for research and development in this domain. Expanding the study to include a broader range of Dark Web marketplaces would provide a more comprehensive dataset and deeper insights into LLM performance across different structures. This expansion would also facilitate further experimentation with new prompts, models, and evaluation strategies.

Further work is also required to improve entity extraction capabilities. Extracting more nuanced entities, such as vendor information, user feedback, or geographical data, would greatly enhance the value of the dataset. This would likely involve fine-tuning LLMs for specific tasks and possibly integrating them with other specialized NLP tools.

Addressing the ethical filtering in LLMs will be crucial for future work. The models' tendency to avoid extracting sensitive content, like weapons listings, due to ethical guidelines limits data comprehensiveness. Future efforts should focus on adjusting prompts or fine-tuning models to overcome these constraints and ensure important information is captured.

Additionally, the creation of a named entity recognition (NER) dataset from the pipeline developed in this study would

represent a significant contribution to the field. Such a dataset could be used to train a NER model specifically designed for Dark Web marketplaces, potentially improving the accuracy and consistency of entity extraction. This, in turn, would help to reduce the reliance on LLMs and provide a more targeted approach to structured information extraction.

Finally, leveraging the extensive knowledge embedded within LLMs could enhance future results. By incorporating external knowledge into the models, it may be possible to improve their ability to interpret and extract more complex information, leading to more accurate and insightful analyses. This approach could involve hybrid models that combine LLMs with domain-specific knowledge bases to further refine the extraction process.

It's worth noting that this research contributes to the broader discourse on countermeasures for cybercrime, with the aim of ensuring the safety and security of online spaces. This work facilitates law enforcement in responding efficiently by expediting the identification of potential threats and unlawful activities on the Internet. All in all, the research offers an important step in the fight against the illegal sale of firearms on online forums. Further progress in this critical domain involves embracing emerging technologies and strengthening efforts to protect communities from potential threats.

ACKNOWLEDGMENT

This work was framed in the context of the project Ceasefire, which receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement 101073876.

REFERENCES

[1] R. Montasari and A. Boon, "An analysis of the dark web challenges to digital policing," in *Cybersecurity in the Age of Smart Societies: Proceedings of the 14th International Conference on Global Security, Safety and Sustainability, London, September 2022*, pp. 371–383, Springer, 2023.

[2] D. Chawla, J. Anthony, and L. M. A. Patel, "Unveiling the dark web: An in-depth introduction," *IITM Journal of Information Technology*, pp. 45–54.

[3] D. Laferrière and D. Décary-Hétu, "Examining the uncharted dark web: Trust signalling on single vendor shops," *Deviant Behavior*, vol. 44, no. 1, pp. 37–56, 2023.

[4] C. Heistracher, F. Mignet, and S. Schlarb, "Machine learning techniques for the classification of product descriptions from darknet marketplaces.," in *ICAI*, pp. 128–137, 2020.

[5] B. Alkhatib and R. S. Basheer, "Mining the dark web: A novel approach for placing a dark website under investigation," *International Journal of Modern Education and Computer Science*, vol. 10, no. 10, p. 1, 2019.

[6] M. Ball and R. Broadhurst, "Data capture and analysis of darknet markets," *Available at SSRN 3344936*, 2021.

[7] D. R. Hayes, F. Cappa, and J. Cardon, "A framework for more effective dark web marketplace investigations," *Information*, vol. 9, no. 8, p. 186, 2018.

[8] K. Connolly, A. Klempay, M. McCann, and P. Brenner, "Dark web marketplaces: Data for collaborative threat intelligence," *Digital Threats: Research and Practice*, vol. 4, no. 4, pp. 1–12, 2023.

[9] J. R. Lee, T. J. Holt, and O. Smirnova, "An assessment of the state of firearm sales on the dark web," *Journal of Crime and Justice*, vol. 47, no. 1, pp. 46–60, 2024.

[10] T. J. Holt and J. R. Lee, "A crime script model of dark web firearms purchasing," *American journal of criminal justice*, vol. 48, no. 2, pp. 509–529, 2023.

[11] D. Rhumorbarbe, D. Werner, Q. Gilliéron, L. Staehli, J. Broséus, and Q. Rossy, "Characterising the online weapons trafficking on cryptomarkets," *Forensic science international*, vol. 283, pp. 16–20, 2018.

[12] H. Alyami, M. Faizan, W. Alosaimi, A. Alharbi, A. K. Pandey, M. T. J. Ansari, A. Agrawal, and R. A. Khan, "An ensemble approach to identify firearm listing on tor hidden-services.," *Comput. Syst. Sci. Eng.*, vol. 38, no. 2, pp. 141–149, 2021.

[13] J. K. Saini and D. Bansal, "A comparative study and automated detection of illegal weapon procurement over dark web," *Cybernetics and Systems*, vol. 50, no. 5, pp. 405–416, 2019.

[14] P. Leonidou, N. Salamanos, A. Farao, M. Aspri, and M. Sirivianos, "A qualitative analysis of illicit arms trafficking on darknet marketplaces," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pp. 1–9, 2023.

[15] R. Broadhurst, J. Foye, C. Jiang, and M. Ball, "Illicit firearms and other weapons on darknet markets," *Trends and Issues in Crime and Criminal Justice*, no. 622, pp. 1–20, 2021.

[16] J. Boegershausen, H. Datta, A. Borah, and A. T. Stephen, "Fields of gold: Scraping web data for marketing insights," *Journal of Marketing*, vol. 86, no. 5, pp. 1–20, 2022.

[17] A. A. Sukmandhani, T. Sunjaya, I. P. Saputro, and J. Ohliati, "Data scraping using python for information retrieval on e-commerce with brand keyword," in *2023 8th International Conference on Business and Industrial Research (ICBIR)*, pp. 179–183, IEEE, 2023.

[18] V. Deshmane, P. Musale, P. Joshi, V. Chinta, K. Gokak, I. Dalbhanjan, *et al.*, "Web scraping for e-commerce website," *International Journal for Innovative Engineering & Management Research*, vol. 13, no. 4, 2024.

[19] S. A. Azcoitia, C. Iordanou, and N. Laoutaris, "Measuring the price of data in commercial data marketplaces," in *Proceedings of the 1st International Workshop on Data Economy*, pp. 1–7, 2022.

[20] S. A. A. Shah, M. A. Masood, and A. Yasin, "Dark web: E-commerce information extraction based on name entity recognition using bidirectional-lstm," *IEEE Access*, vol. 10, pp. 99633–99645, 2022.

[21] I. Gur, O. Nachum, Y. Miao, M. Safdari, A. Huang, A. Chowdhery, S. Narang, N. Fiedel, and A. Faust, "Understanding html with large language models. arxiv 2022," *arXiv preprint arXiv:2210.03945*, 2022.

[22] S. Vatsal and H. Dubey, "A survey of prompt engineering methods in large language models for different nlp tasks," *arXiv preprint arXiv:2407.12994*, 2024.

[23] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

[24] S. Zongbin, Z. Liya, and L. Xiaoqiu, "Material information extraction based on local large language model and prompt engineering," *Data Analysis and Knowledge Discovery*, vol. 8, no. 7, pp. 23–31, 2024.

[25] A. Vijayan, "A prompt engineering approach for structured data extraction from unstructured text using conversational llms," in *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, pp. 183–189, 2023.

[26] F. Polat, I. Tiddi, and P. Groth, "Testing prompt engineering methods for knowledge extraction from text," *Semantic Web. Under Review*, 2024.

[27] A. Hikov and L. Murphy, "Information retrieval from textual data: Harnessing large language models, retrieval augmented generation and prompt engineering," *Journal of AI, Robotics & Workplace Automation*, vol. 3, no. 2, pp. 142–150, 2024.

[28] G. Chatzimarkaki, S. Karagiorgou, M. Konidi, D. Alexandrou, T. Bouras, and S. Evangelatos, "Harvesting large textual and multimedia data to detect illegal activities on dark web marketplaces," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 4046–4055, 2023.

[29] L. Richardson, "Beautiful soup documentation," *April*, 2007.

[30] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[31] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[32] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin, "Darkbert: A language model for the dark side of the internet," *arXiv preprint arXiv:2305.08596*, 2023.