Submitted By : Anish Manandhar

Introduction

Large Language Models are the models that are trained on the huge corpus of the dataset , trained to find the patterns in the natural language. It is made up of neural networks and are at the worst that can ever be. The LLMs work with tokenization , embedding and storing into the vector databases for representing the context. The heart of the LLM lies on the transformers. The stochastic maodels use Reinforcement Learning with Human Feedback (RLHF)
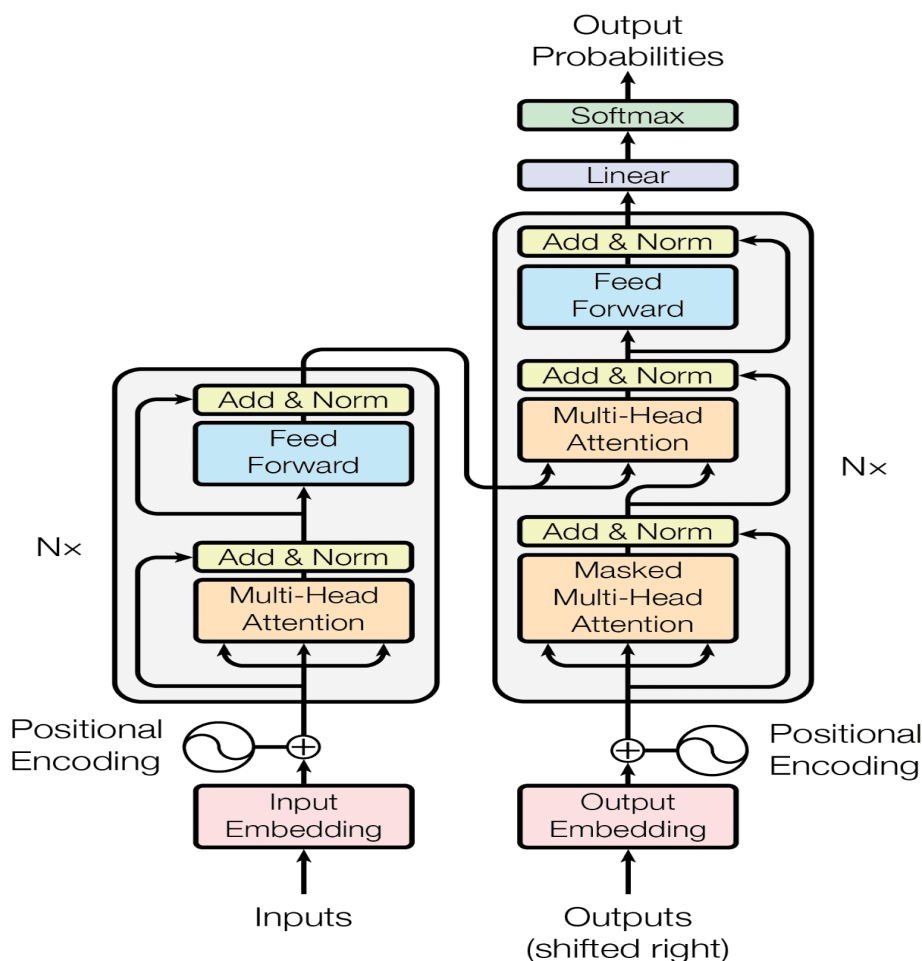

Fig: Transformer

Libraries:

playwright , scrapegraphai==0.9.0b7 ,nest_asyncio, fastapi, uvicorn, transformers, pydentic, requests, beautifulsoup4, --quiet, tiktoken, prettytable, tqdm, nest_asyncio

Problem Statement

The project applies machine learning skills, particularly in natural language processing (NLP), to solve a real-world problem(web-scraping). The main goal is to build an API in a way that extracts meaningful attributes

from an HTML block of an e-commerce website using an open-source large language model (LLM).

Methodology

- **Ollama**

With the help of the Ollama framework, running potent language models locally is simple. Embeddings are a transformative technology in natural language processing, enabling applications to understand and process text semantics at a deep level. By supporting embedding models, Ollama significantly enhances the capabilities of developers and companies to create intelligent and responsive applications. `ollama/nomic-embed-text` is a flexible option for embedding generation because of its exceptional ability to handle both short and long context tasks. This methodology offers developers an additional degree of openness and dependability because it is open source, fully replicable, and open data.

- **Llama3**

Ollama's Llama 3 is a significant update from its predecessor, Llama 2, featuring substantial improvements and new capabilities. Llama 3 has been trained on a dataset seven times larger than Llama 2, and it boasts double the context length at 16,000 tokens. Additionally, it includes a more efficient language encoding system with a larger token vocabulary and exhibits fewer false refusals compared to Llama 2.

For web scraping purposes, Llama 3 can greatly improve textual data extraction, analysis, and comprehension from web sites. Llama 3 may be used to precisely parse complex web material and retrieve pertinent data, like as product details from e-commerce sites or summary points from publications. Llama 3 may also be used for categorizing information, identifying identified entities like as individuals, organizations, and dates, and summarizing lengthy documents. This capability is particularly useful for applications like building databases, extracting contact information, or handling complex queries related to the scraped data.

- **ScrapeGraphAI**
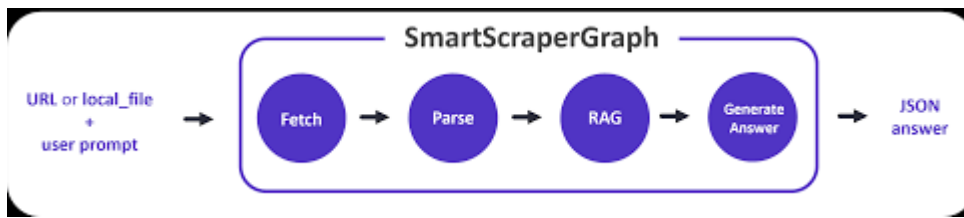
Using LLMs and direct graph logic, ScrapeGraphAI is a potent Python web scraping package that can be used to build scalable and effective scraping pipelines for local documents and websites. This method creates more intelligent and adaptable data extraction methods by fusing the structural efficiency of graph logic with the semantic comprehension powers of LLMs. The llm section specifies the use of the "ollama/llama3" model, setting parameters such as temperature to 0 for deterministic outputs and the format to "json" for structured responses. The base_url is set to "http://localhost:11435", indicating where the LLM is hosted. This allows the library to send HTTP requests to this endpoint, where the LLM processes the input text based on the given prompt and returns a JSON response.

Similarly, the embeddings section configures the use of the "ollama/nomic-embed-text" model, also hosted at "http://localhost:11435". This model converts text into numerical vectors that capture semantic meanings, useful for tasks like similarity searches. The verbose key, set to True, enables detailed logging for debugging purposes, providing insights into the library's operations and interactions with the models.

- Jina AI Jina AI is an open-source project focused on building an ecosystem for neural search. It provides a cloud-native neural search framework to enable developers to build powerful search applications with state-of-the-art AI models. Jina AI offers tools and infrastructure for building scalable and customizable search solutions across various domains, including e-commerce, enterprise search, multimedia search, and more. With Jina AI, users can leverage the latest advancements in deep learning and distributed computing to create intelligent search systems that can understand and retrieve information from large-scale datasets efficiently.

Api Implementation
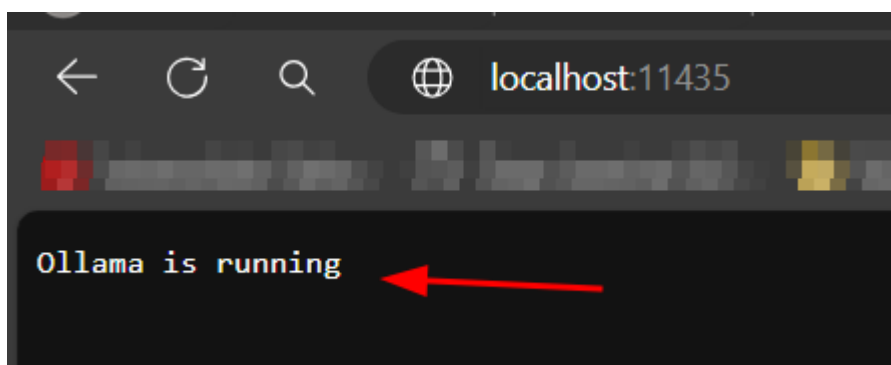

```
graph_config = {
    "llm": {
        "model": "ollama/llama3",
        "temperature": 0,
        "format": "json",
        "base_url": "http://localhost:11435",
```

```
    },
    "embeddings": {
        "model": "ollama/nomic-embed-text",
        "base_url": "http://localhost:11435",
    },
    "verbose": True,
}
```

The graph_config dictionary in the provided code is a configuration setup for LLM and embedding model hosted locally via API endpoints. This configuration allows the library to utilize the models for extracting and processing data from web pages or HTML content.The API implementation involves hosting the specified models locally and making them accessible via HTTP endpoints. These endpoints allow the ScrapeGraphAI library to send HTTP requests to the models for processing tasks. The other implementation includes extracting contents from BeautifulSoup and passing it to JinaAI

In this context, Beautiful Soup helps to fetch and parse HTML which retrieves the HTML content of a web page and parses it into a tree structure. Using its querying capabilities,Jina AI helps to search for specific tags, attributes, and text within the HTML structure. This makes it easier to extract the precise data you need from a web page, such as text within paragraph tags, titles, or any other HTML elements.

Result

Json returned from one of the daraz site

The json retrieved from the LLM is:

```
--- Executing Fetch Node ---
--- Executing Parse Node ---
--- Executing RAG Node ---
--- (updated chunks metadata) ---
--- (tokens compressed and vector stored) ---
--- Executing GenerateAnswer Node ---
Processing chunks: 100%|                                                    | 1/1 [00:00<?, ?it/s]
{'ricants': {'_\ue76c_ Proudly Nepali _\ue76c_ Free Delivery _\ue7f9_ Fashion _\ue7f9_ Beauty _\ue768_ Home & Decor': '', '* TV, Audio / Video, Gaming & Wearables': '', '* Wearable Technology': '', '*
Smartwatches': '', 'Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)': {'Ratings': 6, 'Answered Questions': 11, 'Brand': 'No Brand', 'Price': {
'Rs. 4,000': '', 'Rs. 5,000-20%': ''}, 'Quantity': {'__ __': ''}, 'status': 'Out of stock', 'Add to Wishlist': '', 'Delivery': {'__': ''}, 'Service': {'__ __': ''}, 'Return Policy': {'14 days free & ea
sy return': '', 'Change of mind is not applicable': ''}, 'Warranty': {'6 Months Brand Warranty': ''}, 'Sold by': {'Easy Shop 1581013308': ''}, 'Chat': {'CHAT': ''}, 'Positive Seller Ratings': {'79%': '
'}, 'Ship on Time': {'100%': ''}, 'Chat Response Rate': {'not enough data': ''}, 'Visit Store': '', 'Ratings & Reviews': {'5.0Excellent': 6, 'Sort By': {'Relevance': '', '9***2': '', '3 years ago': '..
.', 'Color Family:Black': ''}}}, 'Product details of Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)': {'* Model: Xiaomi Haylou LS02': '', '*
12 sports modes': '', '* 20 days of endurance': '', '* Intelligent sleep monitoring': '', '* Real-time heart rate monitoring': '', '* Display 1.4 inch TFT Screen': '', '* Battery 260mAh': '', '* Batter
y life: 20 days': '', '* Connectivity Bluetooth V5.0': '', '* Material ABS': '', '* SpecialFeatures Sport modes:': {'Jogging / Fast walking / Biking / Climbing / Spinning / Yoga / Indoor running / Inte
grated training / Gymnastics / Basketball / Football / Rowing': ''}, '* Dustproof and waterproof capability IP68': ''}, 'Specifications of Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monit
oring Call Alert Msg IP68 Waterproof (Black)': {'* Brand': {'No Brand': ''}, '* SKU': {'105587015_NP-1027376026': ''}, '* Material Type': {'Others': ''}, '* Movement': {'Smart': ''}, '* Generation': {'
6th': ''}, '* Watch Features': {'Sleep Monitor': ''}, '* Display Type': {'TFT': ''}, '* Model': {'Ls-02': ''}}, 'What's in the box': {'Charging cable, Smartwatch , User manual': ''}}}
{
  "ricants": {
    "_\ue76c_ Proudly Nepali _\ue76c_ Free Delivery _\ue7f9_ Fashion _\ue7f9_ Beauty _\ue768_ Home & Decor": "",
    "* TV, Audio / Video, Gaming & Wearables": "",
    "* Wearable Technology": "",
    "* Smartwatches": "",
    "Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)": {
      "Ratings": 6,
      "Answered Questions": 11,
      "Brand": "No Brand",
      "Price": {
        "Rs. 4,000": "",
        "Rs. 5,000-20%": ""
      },
      "Quantity": {
```

{'ricants': {'_\ue76c_ Proudly Nepali _\ue76c_ Free Delivery _\ue7f9_ Fashion _\ue7f9_ Beauty _\ue768_ Home & Decor': '', '* TV, Audio / Video, Gaming & Wearables': '', '* Wearable Technology': '', '* Smartwatches': '', 'Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)': {'Ratings': 6, 'Answered Questions': 11, 'Brand': 'No Brand', 'Price': {'Rs. 4,000': '', 'Rs. 5,000-20%': ''}, 'Quantity': {'__ __': ''}, 'Status': 'Out of stock', 'Add to Wishlist': '', 'Delivery': {'__': ''}, 'Service': {'__ __': ''}, 'Return Policy': {'14 days free & easy return': '', 'Change of mind is not applicable': ''}, 'Warranty': {'6 Months Brand Warranty': ''}, 'Sold by': {'Easy Shop 1581013308': ''}, 'Chat': {'CHAT': ''}, 'Positive Seller Ratings': {'79%': ''}, 'Ship on Time': {'100%': ''}, 'Chat Response Rate': {'not enough data': ''}, 'Visit Store': '', 'Ratings & Reviews': {'5.0Excellent': 6, 'Sort By': {'Relevance': '', '9***2': '', '3 years ago': '...', 'Color Family:Black': ''}}}, 'Product details of Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)': {'* Model: Xiaomi Haylou LS02': '', '* 12 sports modes': '', '* 20 days of endurance': '', '* Intelligent sleep monitoring': '', '* Real-time heart rate monitoring': '', '* Display 1.4 inch TFT Screen': '', '* Battery 260mAh': '', '* Battery life: 20 days': '', '* Connectivity Bluetooth V5.0': '', '* Material ABS': '', '* SpecialFeatures Sport modes:': {'Jogging / Fast walking / Biking / Climbing / Spinning / Yoga / Indoor running / Integrated training / Gymnastics / Basketball / Football / Rowing': ''}, '* Dustproof and waterproof capability IP68': ''}, 'Specifications of Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)': {'* Brand': {'No Brand': ''}, '* SKU': {'105587015_NP-1027376026': ''}, '* Material Type': {'Others': ''}, '* Movement': {'Smart': ''}, '* Generation': {'6th': ''}, '* Watch Features': {'Sleep Monitor': ''}, '* Display Type': {'TFT': ''}, '* Model': {'Ls-02': ''}}, 'What's in the box': {'Charging cable, Smartwatch , User manual': ''}}}
{
    "ricants": {
        "_\ue76c_ Proudly Nepali _\ue76c_ Free Delivery _\ue7f9_ Fashion _\ue7f9_ Beauty _\ue768_ Home & Decor": "",
        "* TV, Audio / Video, Gaming & Wearables": "",
        "* Wearable Technology": "",
        "* Smartwatches": "",

```
   "Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call
Alert Msg IP68 Waterproof (Black)": {
     "Ratings": 6,
     "Answered Questions": 11,
     "Brand": "No Brand",
     "Price": {
       "Rs. 4,000": "",
       "Rs. 5,000-20%": ""
     },
     "Quantity": {
       "— —": ""
     },
     "Status": "Out of stock",
     "Add to Wishlist": "",
     "Delivery": {
       "—": ""
     },
     "Service": {
       "— —": ""
     },
     "Return Policy": {
       "14 days free & easy return": "",
       "Change of mind is not applicable": ""
     },
     "Warranty": {
       "6 Months Brand Warranty": ""
     },
     "Sold by": {
       "Easy Shop 1581013308": ""
     },
     "Chat": {
       "CHAT": ""
     },
     "Positive Seller Ratings": {
       "79%": ""
     },
     "Ship on Time": {
       "100%": ""
     },
     "Chat Response Rate": {
       "not enough data": ""
     },
     "Visit Store": "",
     "Ratings & Reviews": {
       "5.0Excellent": 6,
       "Sort By": {
         "Relevance": "",
         "9***2": "",
         "3 years ago": "...",
         "Color Family:Black": ""
       }
     }
   },
   "Product details of Haylou LS-02 Touch Screen Square Shape Smart Watch Heart
```

```
  Monitoring Call Alert Msg IP68 Waterproof (Black)": {
      "* Model: Xiaomi Haylou LS02": "",
      "* 12 sports modes": "",
      "* 20 days of endurance": "",
      "* Intelligent sleep monitoring": "",
      "* Real-time heart rate monitoring": "",
      "* Display 1.4 inch TFT Screen": "",
      "* Battery 260mAh": "",
      "* Battery life: 20 days": "",
      "* Connectivity Bluetooth V5.0": "",
      "* Material ABS": "",
      "* SpecialFeatures Sport modes:": {
        "Jogging / Fast walking / Biking / Climbing / Spinning / Yoga / Indoor
 running / Integrated training / Gymnastics / Basketball / Football / Rowing": ""
      },
      "* Dustproof and waterproof capability IP68": ""
    },
```

Site Name	Beautiful Soup content	Jina AI content
Daraz	` <script>`	Title: Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black) URL Source: https://www.daraz.com.np/products/haylou-ls-02-touch-screen-square-shape-smart-watch-heart-monitoring-call-alert-msg-ip68-waterproof-black-i105587015-s1027376026.html?search=1&spm=a2a0e.searchlistcate
Amazon	`<html> <head> <title>503 - Service Unavailable Error</title> </head> <body bgcolor="#FFFFFF" text="#000000"> <!-- To discuss automated access to Amazon data please contact api-services-support@amazon.com. For information about migrating to our APIs refer to our Marketplace APIs at https://developer.amazonservices.in/ref=rm_5_sv, or our Product Advertising API at https://affiliate-program.amazon.in/gp/advertising /api/detail/main.html/ref=rm_5_ac for advertising use cases. --> <cen`	Title: Amazon.in URL Source: https://www.amazon.in/Memoir-Surgical-Silver-Plated-Stainless/dp/B0771BPK9X/?_encoding=UTF8&content-id =amzn1.sym.0009acdf-e53c-4ef1-bf63-a384e1ba591f:amzn1.symc.cdb151ed-d8fe-485d-b383-800c8b0e3fd3&pd_rd_r=ed7a7a51-162b-4e6e-aa53-7bd4f853ebfe&pd_rd_w=LeHvj&pd_rd_wg=XLEG1&pf_rd_p= 0009acdf-e53c-4ef1-bf63-a384e1ba591f&pf_rd_r=FE2909ZA35203CTX8PPP&ref_=pd_ hp_d_atf_ci_mcx_mr_hp_atf_m Markdown Content: #### Enter the characters you see below Sorry, we just need to make
Flipkart	`<html><body><h1>500 Internal Server Error</h1> An internal server error occurred. </body></html>`	Title: PHILIPS S1151/03 Shaver For Men - PHILIPS : Flipkart.com URL Source: https://www.flipkart.com/philips-s1151-03-shaver-m en/p/itmf70804a8d10f6?fm=organic&iid=239dc8f3-0419-4e31-a6cf-23f0e6241998.SHVGY2SSQYQU47GV.SEARCH&lid=LSTSHVGY2 SSQYQU47GVM471GG&marketplace=FLIPKART&otracker=bro wse&otracker1=hp_rich_navigation_PINNED_neo/mercha ndising_NA_NAV_EXPANDABLE_navigationCard_cc_3_L2_v iew-all&pid=SHVGY2SSQYQU47GV&ppn=browse&ppt=browse &srno=b_1_3&ssid=8t2ig4ekf40000001716740695571&sto re=zlw/79s/

The result shows a flexible framework for scraping content from various websites, calculating associated costs, and presenting the results in a structured manner. By integrating an open-source LLM into the provided codebase, we can develop an API capable of extracting meaningful attributes from HTML content ,even from the link of an e-commerce context, thus addressing the requirements of the assessment task.

Conclusion

However, the open source LLMs are really slow that is why companies prefer OpenAI for the fast scraping Using chatgpt as the API by feeding the HTML contents and processing it:

✎ `<div id="block-pZ-AGMz7zM" class="pdp-block pdp-block__product-detail"><div id="module_flash_sale" class="pdp-block module"><div class="crazy-deal-details pc" style="background-image: url("https://img.alicdn.com/tfs/TB14qayJSf2gK0jSZFPXXXsopXa-620-38.png");"><div class="crazy-deal-details-right"><time class="crazy-deal-details-countdown">Ends in 01:27:01</time><div class="crazy-deal-details-process"><div class="crazy-deal-details-procressbar"><div class="crazy-deal-details-procressbar-inner" style="width: 1%;"></div></div>1 sold</div></div></div></div><div id="module_crazy_deal" class="pdp-block module"></div><div id="module_product_title_1" class="pdp-block module"><div class="pdp-product-title"><div class="pdp-mod-product-badge-wrapper"><img src="https://img.alicdn.com/imgextra/i1/O1CN01cLS4Rj1vgZ8xaij1e_!!6000000006202-2-tps-64-32.png" class="pdp-mod-product-badge" alt="badge" data-spm-anchor-id="a2a0e.pdp.0.i1.7ca73cc7CkGgWM"><span class="pdp-mod-product-badge-title" data-spm-anchor-id="a2a0e.pdp.0.i2 7ca73cc7CkGgWM">X-AGE BUZZ Smart Watch For`
</div> scrape this

🟢 Here is the scraped data from the provided HTML structure:

Flash Sale Details:

- Countdown: Ends in 01:27:01
- Sold: 1 sold

Product Title:

- X-AGE BUZZ Smart Watch For Men and Women (XSW02) - Smart Watch

Product Rating:

- Stars: ★★★★☆
- Total Ratings: 106
- Answered Questions: 26

There can be hallucination after the prompting too:

```
    },
    "* Dustproof and waterproof capability IP68": ""
  },
  "Specifications of Haylou LS-02 Touch Screen Square Shape Smart Watch Heart Monitoring Call Alert Msg IP68 Waterproof (Black)": {
    "* Brand": {
      "No Brand": ""  ◄
    },
    "* SKU": {
      "105587015_NP-1027376026": ""
    },
    "* Material Type": {
      "Others": ""
    },
    "* Movement": {
      "Smart": ""
    },
    "* Generation": {
      "6th": ""
    },
    "* Watch Features": {
      "Sleep Monitor": ""
    },
    "* Display Type": {
      "TFT": ""
    },
    "* Model": {
      "Ls-02": ""
    }
```

Fig : Hallucination with the LLM

Refences

1. https://medium.com/@amanatulla1606/llm-web-scraping-with-scrapegraphai-a-breakthrough-in-data-extraction-d6596b282b4d
2. https://www.youtube.com/watch?v=2BTI3KIiGHU