Research Report on Web Knowledge Extraction

Isaac Zheng

1. Introduction

The rapid evolution of artificial intelligence (AI) technology has unlocked new opportunities across various disciplines. However, despite many strides forward, current AI systems present a significant challenge: generating accurate search results from web-based data.

This challenge can be divided into two primary aspects: enhancing single-hop performance, which involves analyzing a single document or webpage to answer a given question, and improving multi-hop performance, where the AI model must navigate through multiple documents or webpages to answer a question, determining which information source to explore next. Presently, advanced models like GPT-4 have demonstrated impressive single-hop performance, surpassing smaller models even when fine-tuned for specific tasks. However, when it comes to tasks requiring reasoning abilities, even GPT-4 exhibits inconsistencies and inaccuracies. This issue stems from the limitations imposed by the structures of current large language models. While transformer-based models strive to maximize the probability of finding an answer, they lack true reasoning capabilities.

Therefore, the primary focus of this research is to enhance the precision of one-hop question answering (QA) and AI data retrieval and interpretation from the web. The objective is to conceive, develop, and test methodologies that can substantially amplify the accuracy of AI outputs.

2. My Work

In order to enhance the precision of one-hop question answering (QA) and AI data retrieval and interpretation from the web, a tree-of-thoughts structure was employed in the development of the program. The program begins by performing several initial checks to ensure grammatical correctness, question format, the presence of mathematical operations, and whether the answer would be in the format of a list. Based on the results of these checks, the program branches into different paths.

One of the main issues addressed in this research is the limitation of token input, coupled with the diminished performance on sizable input observed in some LLMs. To circumvent such restrictions, I used the common strategy where content is segmented into multiple portions, which are then processed individually.

However, conventional methods for dividing content into blocks of a fixed length have their shortcomings. Notably, vital contextual information is often lost when content is arbitrarily cleaved - an impediment to model outputs seeking to maintain accuracy and coherence. To counteract this issue, I have developed an algorithm designed to prevent any HTML elements from being cut in half unless absolutely necessary. An additional feature of this algorithm is an input stride allowing users to set the overlap between two chunks, thus minimizing the risk of losing context or incorrectly segmenting the information.

The second challenge involves the conversion of HTML to human-readable language whilst maintaining structural information. Although seemingly complex, the solution manifests in a relatively straightforward manner due to existing precedents. The HTML can be transformed into markdown, employing indentation representations for HTML structural information.

However, LLMs such as GPT frequently err in processing HTML, particularly with repetitively structured data, for example, pairs of names and contact information. A pragmatic workaround, following experimentation, is the inclusion of dividers (e.g., multiple underscores) to demarcate each data group. Yet, programmatically ensuring correct divider placement presents another layer of complexity.

To maneuver this conundrum, I have utilized a metric to ascertain groups of elements following the same pattern, hereafter named "lists." Factors considered in this process include elements sharing an identical structure, elements with matching tag names and at least one similar classname, and sequences of recurring elements sharing tag names and classnames.

These "lists" of pattern-similar elements, irrespective of duplication, are subsequently fed into the GPT, the outputs of which are consolidated. Despite this approach, hallucinations of LLM remains an issue. Through experimentation, it was observed that GPT-4, being the most advanced model to date, minimizes such errors on small inputs, provided the input is in an optimal format.

In the final program, the initial check determines whether the query involves summarizing a web page. If this is not the case, the web page is fragmented into smaller chunks. The outputs from these smaller chunks, alongside the "lists," provide the resultant integrated solution.

3. Experiments

Several factors hindered the ability to conduct large-scale experiments for this research. The following reasons contributed to the constraints:

One significant limitation was the cost associated with utilizing GPT-4 for experimentation purposes. Given the substantial computational resources required by this model, the expenses incurred proved to be prohibitive for conducting extensive experiments at a larger scale.

The program's current runtime performance presented another constraint. While striving to improve accuracy, certain trade-offs were made to optimize runtime efficiency. As a result, the program is currently slower than desired, which further constrained the scale of experiments that could be conducted within the available timeframe.

Another challenge encountered was the absence of a suitable dataset for evaluating hard one-hop questions that require advanced reasoning abilities. The existing datasets predominantly encompass simpler questions, lacking the necessary complexity and reasoning demands. As this research was conducted during a limited summer period, it was not feasible to create a comprehensive dataset that would adequately address the research problem.

Despite the limitations, GPT-4 was selected as the model for this research. The general nature of the research problem made it exceedingly challenging for other models, including those with fine-tuning capabilities, to surpass the performance of GPT-4.

As an illustration of the program's capabilities, an experiment was conducted using the question, "List all the academic departments at UIUC." By analyzing a single webpage from the link provided (https://directory.illinois.edu/departmentsUnits), the program generated a response that significantly outperformed existing alternatives such as Bing Chat, Bard, and other web-accessible chatbots, models, or agents. The program's output demonstrated superior accuracy and efficacy in retrieving the desired information from the web.

The academic departments at the University of Illinois at Urbana-Champaign include:
Agricultural and Biological Engineering, Department of
Agricultural and Consumer Economics, Department of
Air Force Aerospace Studies, Department of
Animal Sciences, Department of
Anthropology, Department of
Architecture, School of
Art and Design, School of
Art History Program
Campus Recreation, Department of
Curriculum and Instruction, Department of
Dance, Department of
Environmental Engineering and Science
Evolution, Ecology, and Behavior, Department of
Integrative Biology, School of
Intercollegiate Athletics, Division of
Interdisciplinary Health Sciences (i-Health)
Interdisciplinary Health Sciences Institute (IHSI)
Journalism, Department of
Kinesiology and Community Health, Department of
Labor and Employment Relations, School of

Landscape Architecture, Department of
Latin American and Caribbean Studies, Center for
Latina/Latino Studies, Department of
Law, College of
Liberal Arts & Sciences, College of
Maize Genetics Cooperation/Stock Center
Materials Research Laboratory
Media, College of
Medieval Studies, Program in
Micro and Nanotechnology Lab
Music, School of
National and International Scholarship Program
National Great Rivers Research and Education Center
National Soybean Research Laboratory (ACES, College of)
Native American House
Natural Resources and Environmental Sciences, Department of
Naval Science, Department of
NCSA - National Center for Supercomputing Applications
NESSIE (Net-driven Employee Self-Service and Information Environment)
Neuroscience Program
Nuclear, Plasma, and Radiological Engineering, Department of
Nutritional Sciences Interdisciplinary Graduate Program, Division of
OBFS - Enterprise Risk Management
OBFS - Purchasing
Office for Mathematics, Science, and Technology Education (MSTE)
Office of Civic Life
Office of Student Financial Aid-Administration
Office of the Chief Information Officer
Office of Threat Assessment
Office of Undergrad Research
OLLI - Osher Lifelong Learning Institute
Online Master of Science Teaching Biology Program
Online Programs
Online, UI
Organizational Effectiveness
Organizational Research, Office of (College of Business)
Department of Accountancy
Department of Advertising
Department of Aerospace Engineering
Department of African American Studies
Department of Agricultural and Biological Engineering
Department of Agricultural and Consumer Economics
Asian American Studies, Department of
Astronomy, Department of
Atmospheric Sciences, Department of
Biochemistry, Department of
Bioengineering, Department of
Business Administration, Department of
Chemical and Biomolecular Engineering, Department of
Chemical Sciences, School of
Chemistry, Department of
Civil and Environmental Engineering, Department of
Classics, Department of the
Communication, Department of
Comparative Biosciences, Department of
Computer Science, Department of
Cooperative Extension Service, U of I
Coordinated Science Lab

Council of Academic Professionals (CAP)
Council on Teacher Education
Unit for Criticism and Interpretive Theory
Department of Crop Sciences
Department of Curriculum and Instruction
Department of Dance
Disability Research Institute
Division of Disability Resources and Educational Services
Diversity & Social Justice Education
DMI - Division of Management Information
Early Childhood Research and Practice
East Asian and Pacific Studies, Center for
East Asian Exchange Programs (Institutional and Faculty International Collaboration)
East Asian Languages and Cultures, Department of
Ecology, Evolution and Conservation Biology, Program in
Economic and Business Research, Bureau of
Economic Education, Center for (Economics, Department of)
Education Policy, Organization and Leadership
Education, College of
Educational Psychology, Department of
eLearning, Gies College of Business
Electrical and Computer Engineering, Department of
Energy and Sustainability Engineering
Engineering, Grainger College of
English, Department of
Entomology, Department of
French & Italian, Department of
Gender and Women's Studies, Department of
General Studies, Division of
Geography & Geographic Information Science, Department of
Geology, Department of
Germanic Languages and Literatures, Department of
Global Studies - CGS, Center for
Government and Public Affairs - IGPA, Institute of
Graduate College
Greek Studies, Modern
Plant Biology, Department of
History, Department of
Human Development & Family Studies, Department of
Crop Sciences, Department of
Linguistics, Department of
Literatures, Cultures and Linguistics, School of
Literatures, Cultures, Linguistics Courses, School of
Manufacturing Engineering
Materials Science and Engineering, Department of
Mathematics, Department of
Mechanical Science and Engineering, Department of
Media and Cinema Studies, Department of
Microbiology, Department of
Military Science (Army ROTC), Department of
Molecular and Cellular Biology, School of
Molecular and Integrative Physiology, Department of
College of Business: Organizational Research, Office of
Department of Pathobiology
Department of Philosophy
Department of Physics
Department of Plant Biology
Department of Political Science

Department of Psychology
Slavic Languages and Literatures, Department of
Social Work, School of
Sociology, Department of
Special Education, Department of
Speech and Hearing Science, Department of
Statistics, Department of
Spanish and Portuguese, Department of

4. Future Work

4.1 Enhancing One-Hop Accuracy

The current program shows promise in improving one-hop accuracy; however, there are several avenues for further enhancement. Firstly, to explore alternative models, it is worth considering the integration of open-source models with lower overall performance compared to GPT-4. By experimenting with different models, we can potentially identify a more suitable fit for the specific requirements of one-hop question answering and AI data retrieval tasks. Additionally, introducing parallelism to the program can provide opportunities for optimizing its performance at most stages.

Another promising approach is to create an advanced knowledge graph by combining knowledge graph techniques, textual embedding methods, and large language models. This integrated knowledge graph can serve as a valuable resource for both search and assistance to language models like GPT-4. Leveraging the combined power of these techniques has the potential to significantly improve accuracy in generating precise and relevant results.

Furthermore, it is worth exploring the use of intelligent agents that can interact directly with webpages, rather than relying solely on parsing HTML and identifying useful hyperlinks.

Oftentimes, critical information is not readily available in the initial HTML content and may require interaction, such as clicking buttons or expanding sections on the webpage. By developing agents capable of these interactions, we can increase the chances of retrieving comprehensive and accurate data.

4.2 Improving Multi-Hop Accuracy

Improving multi-hop accuracy is another area that shows promise for future work. Several techniques have already demonstrated positive results in this regard. For instance, fine-tuning a model to predict the next relevant piece of text has proven effective in guiding the exploration of multiple documents. Additionally, incorporating similarity measures, such as Euclidean distance, between the generated text and the embeddings of each document can aid in determining the most relevant document to explore next.

There is also potential for leveraging graph-based approaches to accelerate and enhance the accuracy of the multi-hop process. By utilizing graph structures, we can represent the relationships between documents and use efficient algorithms to navigate and retrieve relevant information. This approach has the potential to further refine the multi-hop performance of the program.

In conclusion, future work should focus on exploring alternative models, enhancing parallelism, leveraging advanced knowledge graphs, developing intelligent agents, and incorporating graph-based techniques to elevate the accuracy and efficiency of both one-hop and multi-hop question answering and data retrieval from the web. These efforts will contribute to the ongoing progress in overcoming the challenges posed by current AI systems and unlock new possibilities for accurate and comprehensive information retrieval.