International Journal of All Research Education and Scientific Methods (IJARESM)
ISSN: XXXX-XXXX, Volume No., Issue No, Month 2015

JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

WEBpedia

(Creating Wikipedia-like Articles from Scratch Using Large Language Models)

[1]Anurodh Pancholi, [2]Priyaanshu Singh

Student School of Computing Science and Engineering
Specialization in Artificial Intelligence & Machine Learning
VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore, Madhya Pradesh - 466114

WEBpedia

1 Dr. M.K. Jayanthi Kannan, 2Anurodh Pancholi, 3Priyaanshu Singh
1 Professor, 2 Student School of Computing Science and Engineering,
3 Student School of Computing Science and Engineering
VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore, Madhya Pradesh - 466114

International Journal of All Research Education and Scientific Methods (IJARESM)
ISSN: XXXX-XXXX, Volume No., Issue No, Month 2015

Abstract

We are going to explore the application of large language models for crafting comprehensive and well-structured long-form articles from scratch, aiming to achieve a level of breadth and depth comparable to Wikipedia entries. This relatively unexplored task presents unique challenges, particularly in the pre-writing phase, such as how to effectively research the topic and develop a detailed outline before beginning the writing process. To address these challenges, we are going to introduce WEBpedia, a writing framework designed for the Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking. WEBpedia will enhance the pre-writing process by (1) uncovering diverse viewpoints during the research phase, (2) simulating discussions where writers with varying perspectives pose questions to a topic expert, leveraging reliable online sources, and (3) organizing the gathered information into a coherent outline.

For evaluation, we will be introducing UtliedWIKI, a dataset comprising recent high-quality Wikipedia articles, and establish outline assessment criteria to evaluate the pre-writing process. Additionally, we will also collect the feedback from Wikipedia editors through search and APIs. Compared to articles that are generated by a traditional outline-driven retrieval-augmented baseline, we aim that WEBpedia articles are going to be found to be more organized and broader in scope.

I. INTRODUCTION

Large language models (LLMs) have shown remarkable writing abilities, yet their potential for crafting grounded, long-form articles, such as comprehensive Wikipedia pages, remains uncertain. Creating such expository content, aimed at informing readers in a well-organized manner, necessitates extensive research and careful planning in the pre-writing phase, even before the actual writing begins. However, previous work on generating Wikipedia articles has often overlooked this crucial pre-writing stage. For example, Liu et al. assume that reference documents are already provided, while Fan and Gardent take for granted that an article outline is available, concentrating instead on expanding each section. These assumptions are not universally applicable, as gathering references and developing outlines require advanced information literacy skills to effectively identify, evaluate, and organize external sources—a challenge even for seasoned writers. Automating this process could help individuals embark on in-depth learning about a topic and significantly reduce the costly expert hours typically needed for expository writing.

We address these challenges by exploring the process of generating Wikipedia-like articles from scratch. We break this problem down into two key tasks. The first task involves conducting research to create an outline, which includes a list of multi-level sections, and gathering a collection of reference documents. The second task uses this outline and the references to write the complete article. This task decomposition reflects the typical human writing process, which generally includes the stages of pre-writing, drafting, and revising.We address these challenges by focusing on generating Wikipedia-like articles from scratch, breaking the task into two key phases. The first phase involves conducting research to create an outline—a structured list of multi-level sections—and gathering a set of reference documents. The second phase uses this outline and references to craft the full-length article. This task decomposition reflects the human writing process, which typically includes pre-writing, drafting, and revising stages (Rohman, 1965; Munoz-Luna, 2015).

While pre-trained language models hold substantial parametric knowledge, a straightforward approach would be to rely solely on this knowledge to generate outlines or complete articles (Direct Gen). However, this method often lacks detail and introduces hallucinations (Xu et al., 2023), especially when dealing with niche or long-tail topics (Kandpal et al., 2023). This limitation highlights the necessity of incorporating external sources. Current approaches frequently employ retrieval-augmented generation (RAG), but these methods still face challenges in the research phase, as simple topic searches often fail to surface critical information.

The design of WEBPEDIA is grounded in two key hypotheses: (1) diverse perspectives inspire varied questions, and (2) crafting in-depth questions requires iterative research. Building on these principles, WEBPEDIA adopts a unique multi-stage approach. It begins by uncovering diverse perspectives through the retrieval and analysis of Wikipedia articles on related topics, then embodies the LLM with distinct viewpoints for generating questions .To facilitate iterative research and generate follow-up questions , WEBPEDIA simulates multi-turn conversations, anchoring the answers to these questions with information sourced from the Internet. Finally, leveraging both the LLM's internal knowledge and the

gathered information, WEBPEDIA constructs an outline that can be progressively expanded to create a comprehensive Wikipedia-like article.

Our main contributions include:

• To evaluate the capacity of LLM systems at generating long-form grounded articles from scratch, and the pre-writing challenge in particular, we curate the UtliedWIKI dataset and establish evaluation criteria for both outline and final article quality.

• We propose WEBPEDIA, a novel system that automates the pre-writing stage. WEBPEDIA researches the topic and creates an outline by using LLMs to ask incisive questions and retrieving trusted information from the Internet.

• Both automatic and human evaluation demonstrate the effectiveness of our approach. Expert feedback further reveals new challenges in generating grounded long-form articles.

II. LITERATURE REVIEW

Table 1: Literature Review of WEBpedia

S No.	Study	Objective	Techniques Used
1	WEBBRAIN: LEARNING TO GENERATE FACTUALLY CORRECT ARTICLES FOR QUERIES BY GROUNDING ON LARGE WEB CORPUS	FUNCTION OF THE MODULE IS TO COLLECT DATA AND DO THE DATA CLEANING PROCESS OF THE GIVEN DATA.	DEEP LEARNING MODELS, INCLUDING TRANSFORMER ARCHITECTURES. NATURAL LANGUAGE PROCESSING (NLP), FOCUSED ON FACT-CHECKING AND QUERY- BASED CONTENT GENERATION.
2	TEACHING LANGUAGE MODELS TO SUPPORT ANSWERS WITH VERIFIED QUOTES	RECENT LARGE LANGUAGE MODELS OFTEN ANSWER FACTUAL QUESTIONS CORRECTLY. BUT USERS CAN'T TRUST ANY GIVEN CLAIM A MODEL MAKES WITHOUT FACT-CHECKING, BECAUSE LANGUAGE MODELS CAN HALLUCINATE CONVINCING NONSENSE.	REINFORCEMENT LEARNING FROM HUMAN PREFERENCES (RLHP): USED TO TRAIN MODELS THAT GENERATE ANSWERS BASED ON SPECIFIC EVIDENCE. SEARCH ENGINES: TO RETRIEVE SUPPORTING EVIDENCE FROM MULTIPLE DOCUMENTS. OPEN-BOOK QA MODELS: THESE MODELS GENERATE ANSWERS WHILE CITING EVIDENCE.
3	AUTOMATICALLY GENERATING WIKIPEDIA ARTICLES: A STRUCTURE-AWARE APPROACH	WE INVESTIGATE AN APPROACH FOR CREATING A COMPREHENSIVE TEXTUAL OVERVIEW OF A SUBJECT COMPOSED OF INFORMATION DRAWN FROM THE INTERNET. WE USE THE HIGH-LEVEL STRUCTURE OF HUMAN AUTHORED TEXTS TO AUTOMATICALLY INDUCE A DOMAIN SPECIFIC TEMPLATE FOR THE TOPIC STRUCTURE OF A NEW OVERVIEW.	MULTI-DOCUMENT SUMMARIZATION TECHNIQUES: USED FOR EXTRACTING AND SUMMARIZING INFORMATION FROM VARIOUS SOURCES. DOMAIN-SPECIFIC TEMPLATES: AUTOMATICALLY GENERATED BASED ON HUMAN-AUTHORED TEXT STRUCTURES FOR CONTENT ORGANIZATION (E.G., MEDICAL ARTICLES). CONTENT SELECTION ALGORITHMS: A METHOD TO EXTRACT RELEVANT CONTENT USING TOPIC-SPECIFIC EXTRACTORS.
4	A CRITICAL EVALUATION OF EVALUATIONS FOR	HUMAN EVALUATION ANALYSIS MODULE: ANALYZES HUMAN	LONG-FORM QUESTION ANSWERING (LFQA): FOR GENERATING AND EVALUATING COMPREHENSIVE

	LONG-FORM QUESTION ANSWERING	EVALUATIONS TO IDENTIFY NEW EVALUATION ASPECTS LIKE COMPLETENESS AND COHERENCE. AUTOMATIC TEXT	ANSWERS.HUMAN EVALUATION METHODS: INVOLVING DOMAIN EXPERTS FOR PREFERENCE JUDGMENTS AND JUSTIFICATIONS.
5	IMPROVING LONG STORY COHERENCE WITH DETAILED OUTLINE CONTROL	DETAILED OUTLINER MODULE: GENERATES A DETAILED, HIERARCHICAL STRUCTURE FOR THE STORY OUTLINE, EASING THE CREATIVE BURDEN DURING DRAFTING. DETAILED CONTROLLER MODULE: ENSURES THE GENERATED STORY ALIGNS WITH THE STRUCTURED OUTLINE BY CONTROLLING HOW PASSAGES FOLLOW THE OUTLINED DETAILS.	STORY GENERATION ALGORITHMS: FOR AUTOMATICALLY CREATING COHERENT LONG-FORM NARRATIVES. EVALUATION METRICS: FOR ASSESSING PLOT COHERENCE, RELEVANCE, AND INTERESTINGNESS BASED ON HUMAN FEEDBACK.
6	ATTRIBUTED QUESTION ANSWERING: EVALUATION AND MODELLING FOR ATTRIBUTED LARGE LANGUAGE MODELS	LANGUAGE MODEL MODULE: A CORE LLM RESPONSIBLE FOR GENERATING ANSWERS IN RESPONSE TO QUERIES. ATTRIBUTION MODULE: RESPONSIBLE FOR TRACKING AND LINKING THE GENERATED TEXT TO ITS SOURCE OR EVIDENCE. EVALUATION FRAMEWORK: A BENCHMARKING SYSTEM USING HUMAN ANNOTATIONS AND AUTOMATED METRICS FOR MEASURING ATTRIBUTION ACCURACY.	LARGE LANGUAGE MODELS (LLMS): UNSUPERVISED MODELS USED FOR GENERATING AND PROCESSING TEXT. ATTRIBUTED QA (QUESTION ANSWERING): A TASK FOCUSED ON LINKING GENERATED TEXT TO RELIABLE SOURCES.

The literature on generating Wikipedia-like articles using large language models reveals a clear progression in addressing the complexities of grounded, long-form content creation. While earlier approaches, such as WikiSum and template-based methods, relied heavily on pre-supplied references or outlines, these methods fell short in simulating the comprehensive pre-writing process essential for high-quality expository writing. Modern advancements in LLMs demonstrate significant potential for generating fluent and coherent content but are hindered by limitations like hallucination, bias, and inadequate handling of niche topics.

The WEBPEDIA framework and UtliedWIKI dataset present innovative solutions to these challenges. By emphasizing pre-writing through perspective-guided question generation, iterative research, and structured outline synthesis, WEBPEDIA addresses the critical gaps in the writing pipeline. UtliedWIKI provides a robust evaluation benchmark, focusing on recent, high-quality articles to mitigate data leakage and assess systems against real-world writing demands.

However, challenges remain, including mitigating source bias, avoiding factual over-association, and expanding multi-modal capabilities. Future research must tackle these issues to refine the reliability and breadth of grounded writing systems further. WEBPEDIA and UtliedWIKI set the stage for these advancements, bridging the gap between human-like writing processes and the capabilities of LLMs in producing comprehensive, structured, and factual articles.
Generating Wikipedia-like articles presents unique challenges, requiring well-organized content, comprehensive research, and reliable source grounding. Traditional approaches often bypass the critical pre-writing phase, assuming the availability of references or outlines. A review of state-of-the-art approaches to Wikipedia article generation highlights several limitations, including reliance on pre-provided references or outlines, lack of iterative research to create multi-level outlines, and evaluation frameworks that prioritize surface metrics like fluency over structural organization and factual grounding. Datasets such as WikiSum (Liu et al., 2018) focus on summarization, lacking emphasis on outline creation or comprehensive article generation. The introduction of datasets like UtliedWIKI, which filters high-quality, recent Wikipedia articles, addresses these gaps by offering a benchmark for grounded, long-form generation.
The WEBPEDIA framework introduces innovations to address these challenges by focusing on the pre-writing stage. It employs perspective-guided question asking to simulate multi-perspective conversations for broader topic coverage and iterative research that grounds multi-turn question-answering in trusted sources. The framework synthesizes this information into structured, multi-level outlines, serving as the foundation for article creation. Complementing this, the UtliedWIKI dataset mitigates training data leakage by curating recent, high-quality Wikipedia articles (B-class or higher) while excluding lists or multi-modal content to prioritize clarity and consistency in evaluations.
To evaluate performance, WEBPEDIA introduces novel metrics inspired by human writing pedagogy (Dietz and Foley, 2019). These include Heading Soft Recall, which measures semantic similarity between generated and human-written outline headings, and Heading Entity Recall, which assesses the coverage of named entities in outlines. Feedback from experienced Wikipedia editors highlights strengths, such as improved organization and breadth of generated outlines, while also identifying challenges like addressing bias in online sources and preventing fabricated connections. By addressing these gaps, WEBPEDIA and UtliedWIKI contribute significant advancements to automated long-form content generation, setting new benchmarks in the field.

III. PROPOSED SYTEM DESIGN

WEBpedia is going to be designed around two key hypotheses: (1) diverse perspectives result in a range of questions, and (2) formulating detailed questions requires iterative research. Building on these ideas, WEBpedia will had a novel multi-stage approach. It will begin by uncovering diverse perspectives through the retrieval and analysis of Wikipedia articles on related topics, then personifies the LLM with specific viewpoints to generate questions. To prompt follow-up questions for iterative research ,WEBpedia will simulate multi-turn conversations where the answers to the generated questions are grounded in information sourced from the Internet. Finally, using the LLM's internal knowledge and the gathered information, WEBpedia constructs an outline that can be expanded section by section to create a comprehensive, Wikipedia-like article.
WEBpedia breaks down generating long articles with citations into two steps:
Pre-writing stage: The system conducts Internet-based research to collect references and generates an outline.
Writing stage: The system uses the outline and references to generate the full-length article with citations.

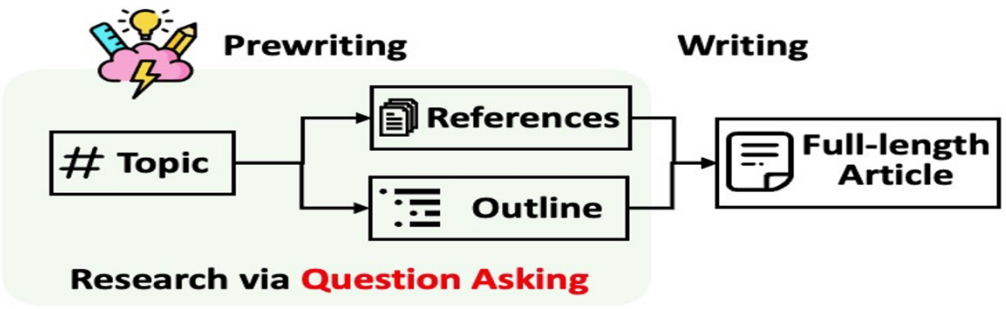

Figure 1

IV. ARCHITECTURE DIAGRAM

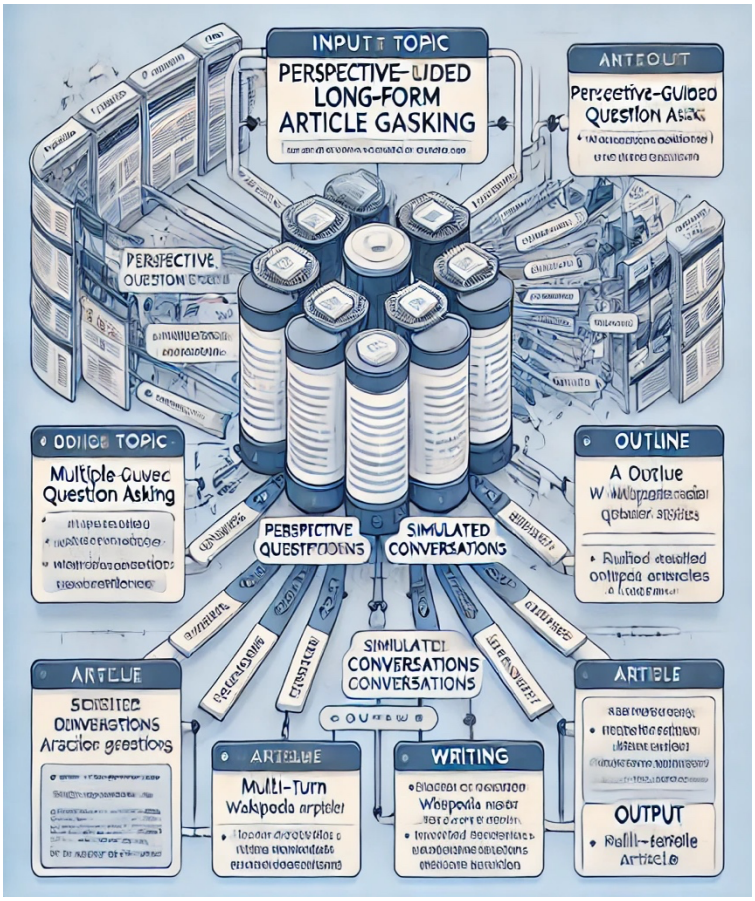

Figure 2: Architecture Diagram

The architecture diagram of the STORM system illustrates how it automates the pre-writing and writing stages of generating grounded, long-form articles. Here's an explanation of each component and its function within the system:

1. Input Topic: The process begins with a given topic. This input is the foundation for the subsequent research and content generation.
2. Perspective-Guided Question Asking:
   - The system first discovers multiple perspectives by surveying Wikipedia articles on related topics.
   - These perspectives guide the question-asking process by prompting the system to generate more focused, in-depth questions about the topic.
   - Each perspective contributes to a more comprehensive understanding of the topic by directing the question-asking from various angles (e.g., basic facts, in-depth technical details, historical context).
3. Simulated Conversations:
   - The system simulates conversations between a Wikipedia writer and an expert.
   - In each conversation round, the LLM generates a question based on its perspective and conversation history, with the expert providing grounded answers based on trusted online sources.

- This conversation process is dynamic, as each answer can generate follow-up questions, iterating to collect more detailed and comprehensive information.

- These conversations help the system gather diverse and detailed information on the topic, which is essential for creating an accurate and well-rounded outline.

4. Curating References:
   - During the simulated conversations, the system retrieves relevant online sources based on the queries and evaluates their trustworthiness.
   - These sources form a reference pool that will later be used to generate the full-length article.
5. Outline Creation:
   - The system creates an initial draft outline ($\mathcal{O}$D) based only on the topic.
   - The outline is then refined using the perspectives and the simulated conversations, resulting in a comprehensive and organized outline ($\mathcal{O}$).
   - The outline serves as the roadmap for the final article and ensures that all necessary sections are covered.
6. Writing the Full-Length Article:
   - Based on the refined outline and gathered references, the system writes the article section by section.
   - For each section, relevant documents from the reference pool are retrieved, and the LLM generates the section content with citations from these sources.
   - The article is generated in parallel for different sections and then concatenated together.
   - The system ensures coherence by eliminating repeated information and synthesizing a lead section summarizing the entire article, in line with Wikipedia's writing style.
7. Final Output:
   - The result is a full-length, grounded Wikipedia-like article that adheres to Wikipedia's structure and writing norms, with a comprehensive coverage of the topic.

The diagram effectively captures the flow of processes from input topic to the final article, showcasing how STORM automates research, question asking, and article generation while maintaining high standards of quality and organization.

V. METHODOLOGY AND ALOGORITHMS USED

The WEBPEDIA framework automates the pre-writing stage of Wikipedia-like article generation through four key processes:

1. Perspective-Guided Question Asking:
   WEBPEDIA identifies diverse perspectives on the topic by analyzing related Wikipedia articles and their tables of contents. Using these perspectives, along with a default "basic facts" viewpoint, the system guides LLMs to generate varied and in-depth questions.
2. Simulating Conversations:
   WEBPEDIA simulates multi-turn conversations between a Wikipedia writer and a topic expert. Questions generated from each perspective are answered with information grounded in trustworthy online sources. Answers are curated by breaking down complex queries into search terms, filtering results according to Wikipedia's reliability guidelines, and synthesizing them for accuracy.
3. Creating the Article Outline:
   A draft outline is initially generated based on the topic using the LLM's intrinsic knowledge. This draft is then refined using insights from the simulated conversations to produce a more comprehensive and detailed outline, which serves as the blueprint for the full article.
4. Writing the Full-Length Article:
   Using the curated references and the outline, WEBPEDIA generates sections of the article in parallel, retrieving relevant information for each section. The sections are combined into a full article, with redundant information removed and a lead section summary synthesized to align with Wikipedia norms. This methodology ensures a structured and well-researched approach to producing long-form articles grounded in reliable information.

We assess WEBpedia using our **UtliedWiki** dataset ,which compiles recent, high- quality Wikipedia articles to prevent data leakage during pre-training. To support the analysis of the pre-writing phase, we establish metrics to evaluate the quality of outlines by comparing them to human-written articles.

VI. METHODOLOGY FOR DEVELOPING WEBpedia

WEBPEDIA identifies the core of automating the research process as automatically coming up with good questions to ask. Directly prompting the language model to ask questions does not work well. To improve the depth and breadth of the questions, WEBPEDIA adopts two strategies:

Perspective-Guided Question Asking: Given the input topic, WEBPEDIA discovers different perspectives by surveying existing articles from similar topics and uses them to control the question-asking process.

Simulated Conversation: WEBPEDIA simulates a conversation between a Wikipedia writer and a topic expert grounded in Internet sources to enable the language model to update its understanding of the topic and ask follow-up questions.

We assess WEBpedia using our **UtliedWiki** dataset ,which compiles recent, high-quality Wikipedia articles to prevent data leakage during pre-training. To support the analysis of the pre-writing phase, we establish metrics to evaluate the quality of outlines by comparing them to human-written articles.

UtliedWiki

Creating a new Wikipedia-like article requires not only fluent writing but also strong research skills. Since modern LLMs are typically trained on Wikipedia text, we will address the potential data leakage by specifically targeting recent Wikipedia articles that were created or significantly edited after the training cutoff of the LLMs we are testing. This approach can also be repeated in the future as new LLMs are developed.

To apply our date-based criteria, we are thinking of focusing on the top 100 most-edited pages each month from February 2023 to august 2024, based on edit counts. To ensure high-quality references, we will filter these articles to include only those rated as B-class or higher by ORES4. We will also exclude list articles and articles without subsections. While high-quality Wikipedia articles often contain structured data (e.g., tables) and are multi-modal, we simplify our task by considering only the plain text component when constructing the dataset.

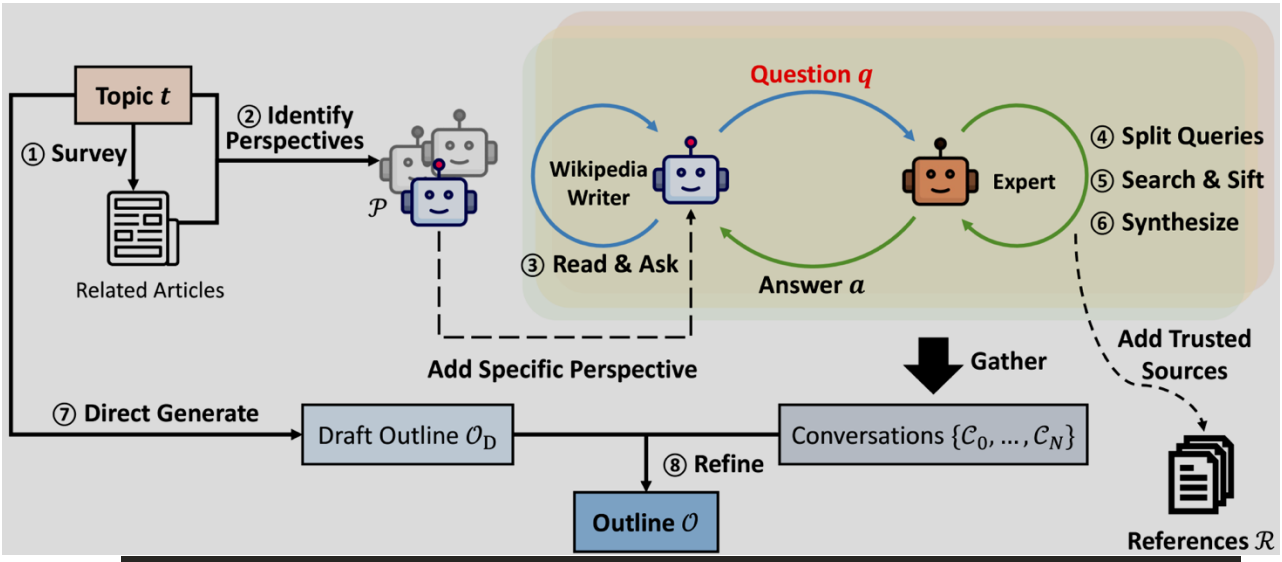

Figure 3: The overview of WEBPEDIA that automates the pre-writing stage. Starting with a given topic, WEBPEDIA identifies various perspectives on covering the topic by surveying related Wikipedia articles. It then simulates conversations between a Wikipedia writer who asks questions guided by the given perspective and an expert grounded on trustworthy online sources. The final outline is curated based on the LLM's intrinsic knowledge and the gathered conversations from different perspectives.

VII. PROJECT FUNCTIONAL MODULES IMPLEMENTATION

- Phase 1: Problem Definition and Research

Define the specific goals, challenges, and requirements for WEBpedia.
Conduct a detailed literature review on large language models, information retrieval, multi-perspective question asking, and automated writing frameworks.
Milestones:
Clear articulation of the problem statement.
Document outlining research gaps and project objectives.

- Phase 2: Data Collection and Preprocessing

Gather data from Wikipedia, online sources, and create the UtliedWIKI dataset.
Preprocess the data to ensure its usability for training the model, including identifying diverse viewpoints and structuring it for retrieval tasks.
Milestones:
Completed dataset creation (UtliedWIKI).
Data preprocessing scripts ready and tested.

- Phase 3: Framework Development

Develop the core components of WEBpedia, including multi-perspective question asking and outline generation.
Implement and integrate information retrieval mechanisms to simulate diverse viewpoints.
Milestones:
Initial version of the WEBpedia framework.
Documentation of algorithms for multi-perspective questioning and outline synthesis.

- Phase 4: Model Training and Tuning

Train the WEBpedia model using UtliedWIKI.
Fine-tune parameters for optimal performance in pre-writing tasks.
Milestones:
Initial model trained and capable of generating detailed outlines.
Hyperparameter tuning completed.

- Phase 5: Evaluation and Testing

Establish the criteria for evaluating the quality of outlines.
Test WEBpedia against baseline retrieval-augmented models to measure improvements in article structure, breadth, and depth.
Collect feedback from Wikipedia editors through APIs.
Milestones:
Evaluation metrics established.
Performance comparison completed.
Feedback from editors collected.

- Phase 6: Final Report and Presentation

Compile findings, results, and improvements into a final project report.
Present the outcomes, with recommendations for future work and potential expansions of WEBpedia.

VIII. LIMITATIONS

In this study, we explore generating Wikipedia-like articles from scratch as a means to advance the field of automated expository writing and long-form content generation. While our approach significantly outperforms baseline methods in both automatic and human evaluations, the quality of machine-generated articles still falls short of well-revised human-

authored counterparts, particularly in terms of neutrality and verifiability. Although WEBPEDIA effectively uncovers diverse perspectives during topic research, the collected information often leans toward dominant online sources and may include promotional content. Furthermore, the identified verifiability challenges extend beyond factual hallucinations, highlighting new obstacles for grounded writing systems.
A further limitation of our work lies in the simplification of the task setup. While we focus on generating Wikipedia-like articles from scratch, our approach is limited to producing free-form text. In contrast, high-quality human-authored Wikipedia articles often integrate structured data and multi-modal content. Addressing the generation of multi-modal, grounded articles remains an area for future exploration.

IX. TECHNOLOGY USED

The report "Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models" highlights various technologies that leverage advancements in natural language processing (NLP), machine learning (ML), and information retrieval to automate the creation of long-form articles. Central to the study are Large Language Models (LLMs), particularly pre-trained architectures like GPT-based models, which are designed to simulate human-like understanding and writing abilities. These models are enhanced through instruction-tuning, a fine-tuning process that leverages instruction-based datasets to improve performance on structured tasks. Key applications of LLMs in the report include simulating multi-perspective conversations for iterative question-asking, generating outlines, organizing content, and crafting full-length, grounded articles.

To address the limitations of LLMs, such as hallucinations and incomplete information, the study integrates Retrieval-Augmented Generation (RAG). RAG combines the parametric knowledge of LLMs with external data sources to enhance factual grounding and reduce errors. Reliable information is sourced from trusted web content and Wikipedia articles, ensuring that generated content is well-supported by references. This integration underscores the importance of information retrieval in grounding answers and providing accurate, verifiable data.

The report also employs Named Entity Recognition (NER) using the FLAIR framework to extract key entities from reference materials and human-written outlines. This step is crucial for evaluating the coverage of generated outlines and ensuring the inclusion of essential concepts from human-authored content. By identifying these entities, the study enhances the quality and relevance of the generated outlines, ensuring alignment with the original material.
Semantic embeddings play a critical role in evaluating machine-generated headings. Using Sentence-BERT, the study calculates heading similarity through cosine similarity, enabling the assessment of semantic alignment between machine-generated and human-written outlines. This ensures that the generated outlines capture the intended structure and meaning of the reference content.
To train and evaluate the models effectively, the report introduces the UtliedWIKI Dataset, a custom dataset curated from recent, high-quality Wikipedia articles. This dataset is carefully selected to mitigate data leakage from pre-training corpora by focusing on articles created or significantly edited after the LLM training cutoff. Additionally, only articles with B-class or higher quality ratings (as determined by ORES) are included, ensuring the dataset's relevance and accuracy.
The evaluation framework in the study includes several innovative metrics. Heading Soft Recall measures the similarity between section headings in machine-generated and human-written outlines using Sentence-BERT embeddings, while Heading Entity Recall assesses the presence of key entities in the generated outlines. Expert feedback from experienced Wikipedia editors further evaluates article quality and identifies challenges such as source bias transfer.

In summary, the report demonstrates how LLMs, when combined with retrieval-based methods, multi-turn simulations, and advanced evaluation techniques, can automate the creation of comprehensive and factually accurate long-form articles. Technologies like FLAIR for NER, Sentence-BERT for semantic similarity, and RAG for knowledge grounding are pivotal in achieving this goal. These innovations showcase the potential of integrating state-of-the-art NLP techniques to tackle complex writing tasks effectively.

X. MOTIVATION & OBJECTIVE

- The project, WEBpedia, aims to address significant gaps and opportunities in automated content generation, particularly for producing long-form, comprehensive, and factually grounded articles like those found on Wikipedia. Its motivations stem from both the practical challenges of article creation and the limitations of existing approaches. One key focus is enhancing the pre-writing process, which includes research, reference gathering, and outline creation—tasks often overlooked or underestimated by current methodologies. Automating these critical steps has the potential to lower barriers for non-experts, democratizing knowledge creation and enabling a broader range of users to generate well-structured, high-quality content.

- Another key motivation is addressing the limitations of large language models (LLMs) in structured planning. While LLMs excel in fluent text generation, they struggle with tasks requiring detailed planning and the synthesis of information from multiple perspectives. WEBpedia seeks to fill this gap by focusing on pre-writing tasks that go beyond basic prompting techniques, allowing LLMs to perform more effectively in research-intensive and organizational roles.

- The project also aims to improve the depth and breadth of knowledge captured in generated content. Wikipedia articles are distinguished by their comprehensive coverage and evidence-backed synthesis, which requires integrating information from diverse perspectives. WEBpedia addresses the challenge of moving beyond surface-level generation to create articles that are both thorough and grounded in reliable references.

- Beyond its research goals, the WEBpedia framework has significant real-world applications in areas like education, journalism, and content marketing. These fields often require users to produce factual, organized, and comprehensive long-form content, even when they lack extensive expertise. By providing tools to streamline this process, the project has the potential to transform how individuals and organizations approach content creation.

- At a broader level, WEBpedia represents a step toward redefining how AI systems handle complex, multi-stage tasks. By emphasizing planning, research, and iterative refinement, the project contributes to ongoing efforts to make AI systems more transparent, reliable, and effective in tasks requiring human-like reasoning and synthesis. This approach underscores the potential of AI to support and enhance human capabilities in producing high-quality, impactful content.

- The objective of the WEBpedia project is multi-faceted, aiming to advance automated long-form content generation through innovative methodologies and robust evaluation frameworks. A central goal is the introduction of the WEBpedia framework, which includes perspective-guided question generation and outline-driven content creation. By leveraging large language models (LLMs), the framework simulates multi-perspective conversations and iterative research processes to uncover diverse viewpoints on a given topic. Additionally, it facilitates the systematic creation of detailed, multi-level outlines to guide the writing process effectively.

- Another key focus is the development and utilization of the UtliedWiki dataset, a curated collection of recent, high-quality Wikipedia articles. This dataset is designed to mitigate data leakage while ensuring relevance in evaluation. It also serves as a foundation for establishing metrics and benchmarks that assess the quality of generated outlines and final articles, allowing for meaningful comparisons with human-written content.

- WEBpedia also addresses real-world challenges in content generation by emphasizing grounded and factually accurate article creation, a critical requirement for Wikipedia-like entries. The project seeks to emulate the human writing process—encompassing pre-writing, drafting, and revising—to enhance coherence and organizational quality in the generated content. This human-like simulation ensures that the output aligns with the rigorous standards of high-quality writing.

- Furthermore, the project aims to establish baselines for future research by highlighting and addressing challenges in generating long-form, grounded content. By comparing WEBpedia's performance against existing systems, the project validates its effectiveness and demonstrates its potential for broader application. This comparison not only underscores the framework's capabilities but also sets the stage for continued research and development in automated content creation.

- By combining innovative frameworks, curated datasets, and comprehensive evaluation criteria, WEBpedia seeks to set a benchmark for utilizing LLMs in producing grounded, long-form articles. The project contributes valuable tools and insights to the field of automated writing systems, fostering advancements in both research and practical applications.

XI. CONTRIBUTION AND FINDINGS

The report makes several key contributions to the field of automated long-form content generation. It introduces WEBPEDIA, a novel framework designed to automate the pre-writing stage of creating Wikipedia-like articles. Unlike traditional methods, WEBPEDIA emphasizes the synthesis of topic outlines through a process of retrieval and multi-perspective question asking. This ensures that the system not only generates a structured outline but also grounds it in reliable information sourced from diverse

perspectives. Additionally, the authors curated the UtliedWIKI dataset, a collection of high-quality, recent Wikipedia articles. This dataset serves as an invaluable resource for evaluating pre-writing and article generation processes while avoiding data leakage from language model pre-training. Furthermore, the report proposes innovative metrics such as heading soft recall and heading entity recall, providing robust tools to assess the quality of generated outlines against human-written counterparts.

Through rigorous evaluation, the study finds that WEBPEDIA significantly improves the organization and breadth of generated articles. Outlines created using the system enhanced the final article's structure and scope, with a 25% increase in organization and a 10% improvement in coverage compared to baseline methods. Incorporating diverse perspectives proved instrumental in creating nuanced outlines that comprehensively address a topic. This approach highlights the importance of accounting for different viewpoints in the research phase, a step often overlooked in traditional content generation models.

Despite its advancements, the study identifies critical challenges in automated long-form article generation. One key issue is source bias transfer, where biases in the retrieved information can influence the neutrality of the content. Another is the over-association of unrelated facts, where language models fabricate connections that do not exist. These challenges underscore the need for further refinement in LLM-based writing systems to ensure accuracy and reliability.

The findings also validate the effectiveness of iterative research methods in automated writing. By simulating multi-turn conversations between a topic expert and a writer, WEBPEDIA enables the generation of detailed and grounded content. This iterative approach ensures that information is refined and enriched over multiple cycles of questioning and answering, resulting in a more comprehensive understanding of the subject matter. While large language models demonstrate impressive capabilities in research and outline creation, the study reveals that issues like hallucination and coherence in long-form generation still necessitate human intervention and further optimization.

Overall, the report's contributions and findings position WEBPEDIA as a transformative tool for automated content creation while identifying areas for future exploration to enhance the reliability and depth of long-form article generation.

XII. CONCLUSION

We are going to introduce WEBpedia, an LLM-based writing system designed to automate the pre-writing stage of creating Wikipedia-like articles from scratch. And To facilitate our study on generating grounded, long-form articles, we will curate a UtliedWiki dataset and establish evaluation criteria. By increasing breadth and depth of the article, WEBpedia helps identify new challenges for grounded writing systems, as highlighted by many research and evaluations.

The WEBPEDIA framework represents a significant advancement in automating the process of writing Wikipedia-like articles from scratch. By focusing on the often- overlooked pre-writing stage, WEBPEDIA emphasizes thorough research and structured content generation, addressing key challenges such as diverse perspective gathering, iterative question asking, and grounded information retrieval. Through its multi-stage approach, WEBPEDIA demonstrates the capability to generate detailed and well-organized outlines, which form the foundation for comprehensive and coherent articles.

Evaluation using the UtliedWIKI dataset and feedback from experienced Wikipedia editors highlight the system's strengths in creating articles with improved breadth, depth, and organization compared to baseline methods. However, challenges such as mitigating source bias and avoiding the fabrication of connections between unrelated facts underscore the need for further refinement in grounded article generation.

Overall, WEBPEDIA showcases the potential of large language models in supporting and enhancing expository writing tasks, particularly for topics requiring meticulous research and planning. This work lays the groundwork for future innovations in long-form content creation, bridging gaps in automation and human-level expertise.

XIII. REFERENCES

[1] **Teaching language models to support answers with verified quotes** (https://arxiv.org/pdf/2203.11147)

[2] **Automatically Generating Wikipedia Articles: A Structure-Aware Approach** (https://aclanthology.org/P09-1024.pdf)

[3] **Attributed Question Answering: Evaluation and Modelling for Attributed Large Language Models** (https://arxiv.org/pdf/2212.08037)

[4] **Improving Long Story Coherence With Detailed Outline Control** (https://aclanthology.org/2023.acl-long.190.pdf)


JETIR