UTS

Research Methodology

in Computer Science

Kelompok 19

Anggota:

1. **Jonathan - 2702253130**
2. **Ricky Herald Cenniago - 2702252714**
3. **Steven Chowina - 2702295373**

---

1. **Research Topic & Problem Formulation**
   a. **Explain the real-world relevance or computing challenge that your group seeks to address.**

      Our group wants to solve the problem in the use of Large Language Models Artificial Intelligence, where Artificial Intelligence tends to generate inaccurate data / hallucinations.

   b. **Present your research problem concisely (2-3 sentences) and discuss how you arrived at your problem statement.**

      The problem we will research is "How to reduce hallucinations in AI LLM when creating research papers?." We chose this issue because we believe it is relevant, especially in the Research Methodology course. In exploring this matter, we also found a solution which is integrating verification algorithms or data post-processing such as STORM to improve the accuracy of AI LLM in generating data.

   c. **Highlight why this problem needs attention in computer science and the potential impact of solving it.**

      According to our group discussion, this issue needs special attention in Computer Science department because a lot of technological developments come from experiments and research, like AI itself. By reducing hallucinations in AI, research paper writing can become easier and grow the technological advancements in research.

2. **Literature Review & Citation Management**
   a. **Provide 6 – 8 critical research papers that directly inform your research problem (closely related papers) using table of comparison. You may add other columns as necessary.**

      Link Spreadsheet: Tabel UTS RM Kelompok 19

No	Paper Title	Research Objective	Data	Method	Result	Future Direction	Link to Paper

1	Assisting in Writing Wikipedia-like Articles from Scratch with Large Language Models	To explore and automate the pre-writing stage of generating comprehensive, grounded Wikipedia-like articles using Large Language Models (LLMs). Specifically, it aims to address the challenges of researching and outlining topics before writing.	FreshWiki, a dataset comprising recent high-quality Wikipedia articles selected to minimize data leakage from LLM training datasets. The dataset includes topics edited after the LLMs' training cutoff.	STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking), which involves (1) discovering multiple perspectives through related Wikipedia articles, (2) simulated conversational questioning grounded in trusted internet sources, and (3) outline refinement and full article generation.	STORM significantly improves article quality, organization (25% better) and breadth (10% better) compared to baseline methods, as evaluated by experienced Wikipedia editors and automated metrics. Editors highlighted improved organization and content coverage as key strengths.	Further enhancement of neutrality, addressing biases from internet sources, improved citation verifiability, expansion into multimodal content generation, and multilingual support.	https://arxiv.org/pdf/2402.14207
2	The state of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value	To survey and analyze the adoption and impact of generative AI (Gen AI) and analytical AI in global organizations, exploring implementation strategies, adoption patterns, associated benefits, and challenges as of early 2024.	Survey data from McKinsey Global Survey on AI with 1,363 respondents across various global regions, industries, and company sizes conducted between February 22 and March 5, 2024.	Quantitative survey analysis providing statistical insights into AI adoption rates, business functions impacted, investment distribution, and perceived risks and benefits. Analysis included comparisons across regions, industries, and job roles.	Significant increase in Gen AI adoption (65% from 33% previous year), substantial business benefits (cost reduction and revenue growth in various business functions), identification of critical risks (particularly inaccuracies), and insights into practices of top-performing organizations.	Emphasizes deeper customization of Gen AI models, improved strategies for risk management (especially inaccuracies and security), advanced governance models for responsible AI usage, and further exploration of hybrid AI strategies (combining off-the-shelf and proprietary/customized solutions).	https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2024/the-state-of-ai-in-early-2024-final.pdf

3	Long Short-Term Memory	To address the vanishing gradient problem in traditional recurrent neural networks (RNNs), enabling effective learning of long-range dependencies over extended sequences.	Artificial datasets involving local, distributed, real-valued, and noisy patterns with time delays.	Introduces the Long Short-Term Memory (LSTM) architecture, using memory cells with constant error flow via "Constant Error Carrousels" (CEC), input/output gating, and multiplicative gate units to control information flow.	LSTM significantly outperforms other recurrent network methods like BPTT, RTRL, Elman nets, and Cascade-Correlation, successfully solving artificial long-time lag tasks that other recurrent algorithms failed to handle.	Further exploration into real-world applications (e.g., speech processing, time series prediction), integrating LSTM into hierarchical sequence models, addressing more complex, multimodal sequence tasks.	https://www.bioinf.jku.at/publications/older/2604.pdf
4	Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation	To develop an RNN-based neural network architecture capable of encoding phrases into fixed-length representations and decoding them into another language, improving phrase-based statistical machine translation (SMT).	English-French translation datasets from WMT'14, including Europarl, news commentary, and crawled corpora totaling hundreds of millions of words.	The RNN Encoder–Decoder model, composed of two recurrent neural networks (encoder and decoder), trained jointly to learn phrase translations by encoding variable-length sequences into fixed-length vectors and decoding these vectors into target sequences.	The RNN Encoder–Decoder significantly improved the BLEU score of phrase-based SMT systems by effectively learning semantically and syntactically meaningful phrase representations.	Further exploration into replacing traditional phrase tables entirely with neural models, applying the approach to other languages, integrating deeper linguistic features, and extending to speech transcription.	https://aclanthology.org/D14-1179.pdf
5	A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions	To systematically review and categorize hallucination phenomena in large language models (LLMs), identifying their causes, detection methods, mitigation techniques, and outlining open research challenges.	Analysis of literature and empirical studies from various academic publications and benchmarks (e.g., TruthfulQA, HaluEval,	Systematic literature review and detailed taxonomy formation, categorizing hallucinations into factuality (factual contradiction and fabrication) and faithfulness (instruction,	Established a clear taxonomy of hallucinations, identified detailed causes at data, training, and inference stages, comprehensively reviewed detection	Research recommended in multimodal hallucinations (vision-language models), improved benchmarks for evaluation, more advanced retrieval and decoding methods, understanding and managing knowledge boundaries, and	https://arxiv.org/abs/2311.05232

			SelfCheckGPT-Wikibio), covering diverse contexts including misinformation and biases.	context, logical inconsistencies); detailed exploration of detection and mitigation strategies including Retrieval-Augmented Generation (RAG).	benchmarks, assessed effectiveness of various mitigation strategies (particularly highlighting challenges and limitations in RAG approaches).	developing stronger alignment methods.	
6	Retrieval-Augmented Generation for Knowledge-Intensive NLP	To develop and evaluate retrieval-augmented generation (RAG) models that combine parametric (pre-trained seq2seq transformer) and non-parametric (dense vector index of Wikipedia) memory systems to improve performance on knowledge-intensive NLP tasks.	Wikipedia (December 2018 dump), divided into approximately 21 million 100-word passages for creating a dense vector index. Evaluated on datasets: Natural Questions (NQ), TriviaQA, WebQuestions, CuratedTrec, MS-MARCO, Jeopardy Question Generation, and FEVER.	RAG models integrate a dense passage retriever (DPR) as non-parametric memory and a pre-trained seq2seq model (BART) as parametric memory, using end-to-end fine-tuning. Two variants: RAG-Sequence (uses same retrieved document for entire generation) and RAG-Token (uses different retrieved documents per token).	Achieved state-of-the-art results in multiple knowledge-intensive NLP tasks (e.g., open-domain QA, abstractive QA, Jeopardy question generation), significantly outperforming baseline methods. Demonstrated more factual, diverse, and accurate generations compared to parametric-only models (BART).	Suggests exploring joint pre-training of retrieval and generation components from scratch, and investigating further integration techniques between parametric and non-parametric memories for various NLP applications.	https://arxiv.org/abs/2005.11401
7	Enabling Large Language Models to Generate Text with Citations	To enable large language models (LLMs) to generate text accompanied by citations, thereby enhancing factual correctness, verifiability, and trustworthiness by providing supporting evidence explicitly within the generated outputs.	Three distinct QA datasets: ASQA, QAMPARI, and ELI5, paired with large retrieval corpora (Wikipedia and Sphere,	Proposes ALCE (Automatic LLM Citation Evaluation), a benchmark requiring models to retrieve relevant passages, generate answers, and properly cite	Demonstrates significant room for improvement, as even top LLMs (ChatGPT, GPT-4) lack proper citation support in ~50% of answers (e.g.,	Future directions include developing improved retrieval methods, advancing models capable of handling longer contexts, and enhancing LLMs' abilities to synthesize accurate information from	https://arxiv.org/abs/2305.14627

			totaling millions of passages) to evaluate citation quality across diverse question types including factoid and explanatory (how/why/what).	supporting evidence. Evaluates performance on three dimensions: fluency (using MAUVE), correctness, and citation quality (using Natural Language Inference - NLI models).	ELI5). Highlights that simpler approaches (e.g., including top retrieved passages in context) can effectively improve correctness and citation quality.	multiple sources simultaneously.	

b. For each paper, explain in paragraphs:

i. How it aligns or conflict with your proposed research direction.

1) We use this paper as reference because it contains the explanation of the STORM framework such as its definition, how it works, experiments, and experiment results analysis.
2) We use this research result as a reference because it contains online survey data on the use of Generative AI.
3) We use this paper as a reference because it contains the contribution of Long-Short Term Memory in the development of AI, especially in the context of how academic research results influence AI technological advancement.
4) We use this paper as a reference because it contains “Transformer Architecture”, which is a crucial architecture in the development of AI.
5) This paper aligns with our research because it also discusses hallucinations, which is the main issue in our research paper.
6) This paper explains the RAG framework, which supports the existence of the STORM framework, our paper’s main discussion.
7) This paper explains the improved RAG framework, where with additional references, AI-generated data will be more accurate and hallucinations minimized, so we use it as a reference because it’s also discussed in STORM.

ii. One limitation or unanswered question in the paper that your study might address.

1) This paper is the main reference source of our research, so we found no limitations or unanswered questions in this paper.
2) This paper only contains statistical results of online surveys regarding the use of Generative AI.
3) This paper discusses Long Short Term Memory which does not address source verification, multi-agent reasoning, or structured planning topics.

4) The limitation in this paper is that algorithms used to develop AI may be flawed, so when built into an LLM, it generates false data and requires methods to reduce this.
5) The limitation in this paper is that while hallucinations are discussed, no solution is provided to reduce them.
6) This paper discusses the RAG framework that functions to reduce hallucinations, but hallucinations still occur so the problem is not fully solved.
7) Same as number 6, this paper discusses improved RAG, but still produces hallucinations so limitations remain.

c. Indicate which citation manager tool you are using and provide the proof (screenshot).

For the citation manager, we do not use a specific tool to manage citations. We use BibTeX, which is LaTeX's native format. We collect references through searching on Google and Google Scholar. Below we show our citations in our LaTeX.

```
\begin{thebibliography}{00}
    \bibitem{b0}Y. Shao et al., "Assisting in Writing Wikipedia-like
    Articles from Scratch with Large Language Models," arXiv:2402.14207
    [cs.CL], 2024.
    \bibitem{b1} McKinsey, "The state of AI in early 2024: Gen AI
    adoption spikes and starts to generate value," 2024.
    \bibitem{b2} S. Hochreiter and J. Schmidhuber, "Long short-term
    memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
    \bibitem{b3} K. Cho et al., "Learning phrase representations using
    RNN encoder-decoder for statistical machine translation," in Proc.
    EMNLP, Doha, Qatar, 2014, pp. 1724-1734.
    \bibitem{b4} L. Huang et al., "A survey on hallucination in large
    language models: principles, taxonomy, challenges, and open
    questions," arXiv:2308.XXXX, 2023.
    \bibitem{b5} P. S. H. Lewis et al., "Retrieval-augmented generation
    for knowledge-intensive NLP," in Advances in Neural Information
    Processing Systems 33, 2020, pp. 9459-9474.
    \bibitem{b6} T. Gao et al., "Enabling large language models to
    generate text with citations," in Proc. EMNLP, Singapore, 2023, pp.
    6465-6488.
\end{thebibliography}
```

3. Research Design & Methodology

a. State whether you plan a descriptive, correlational, experimental, or case study design for your research. Explain how your chosen design aligns with your research objectives. Justify with your findings in literature review.

We are planning to use an experimental design approach in our research with the aim of measuring the effectiveness of the STORM framework in reducing hallucinations in Research Paper generation using AI LLM.

b. State whether you plan a quantitative, qualitative, or mixed-method approach. Explain how your chosen approach will help you generate valid and reproducible findings. Justify with your findings in literature review.

Our group plans to use a Mixed-Method Approach, which we believe is the most suitable for assessing the performance of the STORM framework, where we will use numerical metrics such as citation accuracy and LLM hallucination rates. Additionally, there will also be essay-style assessments regarding the data accuracy generated by AI LLM with STORM.

c. Propose your solution based on your approach and research design. Plan on how you get your expected result.

The solution our group proposes is the implementation of the STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking) framework to help in writing Research Papers with minimal hallucination.

STORM itself works by using various sources to get information, then engages in conversations from different perspectives to compare, and then curates the obtained information to create an Outline.

Research Plan:

- Dataset Collection
- Model Preparation
- Content Collection
- Evaluation
- Data Analytics

With this, we hope to get results showing how research paper writing improves with the use of the STORM framework in reducing hallucinations and improving overall writing quality.

4. Data Handling & Analysis

a. Outline the type of data (structured, unstructured, or both) you expect to gather and explain how you will get them.

The type of data we will use is structured and unstructured data. Both types of data will be used simultaneously, where structured data will provide table views and unstructured data will explain the meaning / relationship between the data and the discussed problem.

The data we will present will be obtained through personal research and questionnaires. The research will be conducted by comparing the results of articles made by AI LLM such as ChatGPT, AI LLM with STORM algorithms such as CO-STORM, and human writers. This comparison will show the quality and accuracy of the data. Besides personal research, we will create a questionnaire that will ask about AI LLM hallucination accuracy and STORM integration in research paper writing.

b. Describe one data cleaning technique you plan to use and the analysis tool (or language) you find most suitable (e.g., Python, R, MATLAB). Explain why these choices fit your project.

The data cleaning technique that our group will use is text normalization. We use this technique because one of the data types we will use is unstructured data, which is articles. Without the normalization process, different texts might discuss the same thing, which would affect the accuracy and hallucination values. This technique will use the Python language, where libraries such as nltk and spaCy will be used for data processing in accordance with the Natural Language Processing course, pandas for managing structured data such as questionnaires, and sklearn for metric analysis and model comparisons.