An Efficient System for Creating Clinical Trial Documents Using LLMs

Dec-9-2024

TIS Inc.
Digital Innovation SBU
Healthcare Services Div. Pharma & Medical Services Dept.
Junya Kamura

Outlines

- About LLMs
  - What is LLMs?
  - Applications of LLMs
  - Advanced LLM s Techniques
  - Points to consider when using large language models
- Clinical Document Generator
  - Case Overview
  - Preparation
  - Implementation
  - Quantitative Evaluation
  - Qualitative Evaluation
  - Output Example
  - Results
  - Future work
- Summary

What is LLMs?

- Large language models are a type of AI model that handles text data and are trained on an extensive amount of text data.
- These models can perform various language-related tasks such as text generation, comprehension, translation, summarization, and question answering.
- Large language models are characterized by their "large-scale" nature in terms of size and the volume of training data.

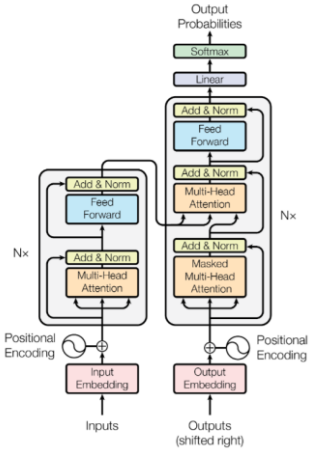
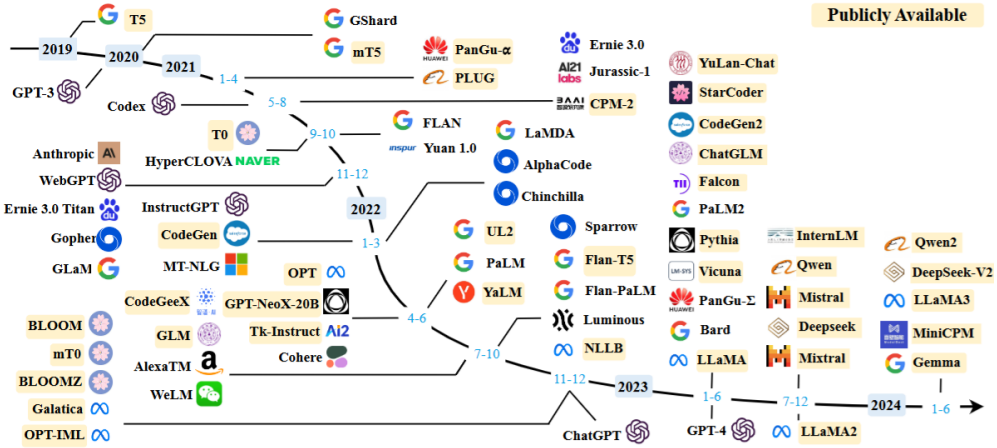


Figure 1: The Transformer - model architecture.



https://arxiv.org/pdf/1706.03762
https://arxiv.org/pdf/2303.18223

Applications of LLMs

Recognition	Generation
Sentiment Analysis	Summarization
Entity Recognition	Text Generation
Text Classification	Text Completion
Question Answering (QA)	Creative Content
Translation	Code Generation

Advanced LLMs Techniques

Few-shot Learning

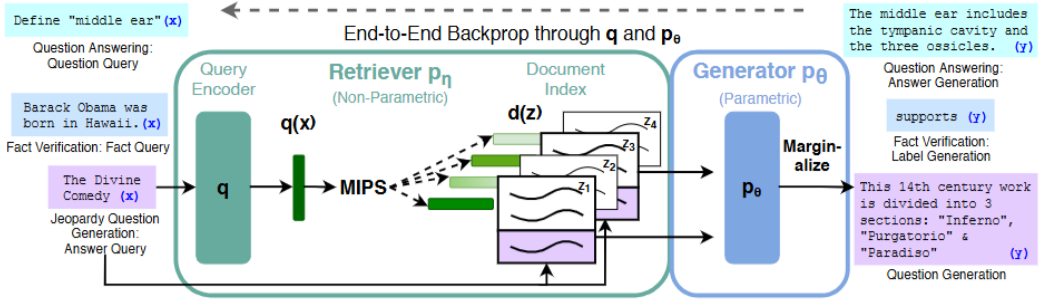
イチロー：オリックスブルーウェーブ
松井秀喜：読売ジャイアンツ
大谷翔平：

大谷翔平：北海道日本ハムファイターズ

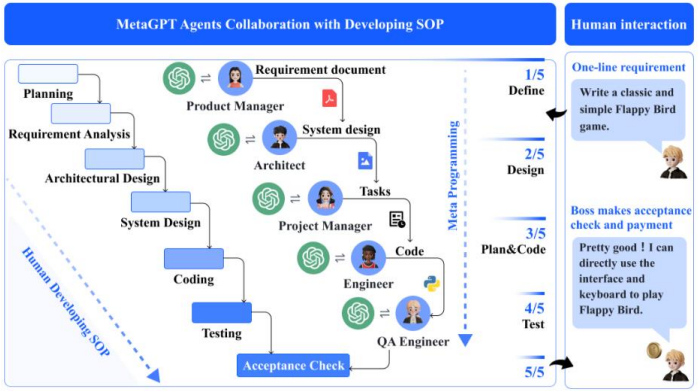
イチロー：マリナーズ
松井秀喜：ヤンキース
大谷翔平：

大谷翔平：エンゼルス

RAG（Retrieval-Augmented Generation）



Multi-Agent



https://arxiv.org/pdf/2005.11401
https://arxiv.org/pdf/2308.00352

Points to consider when using large language models

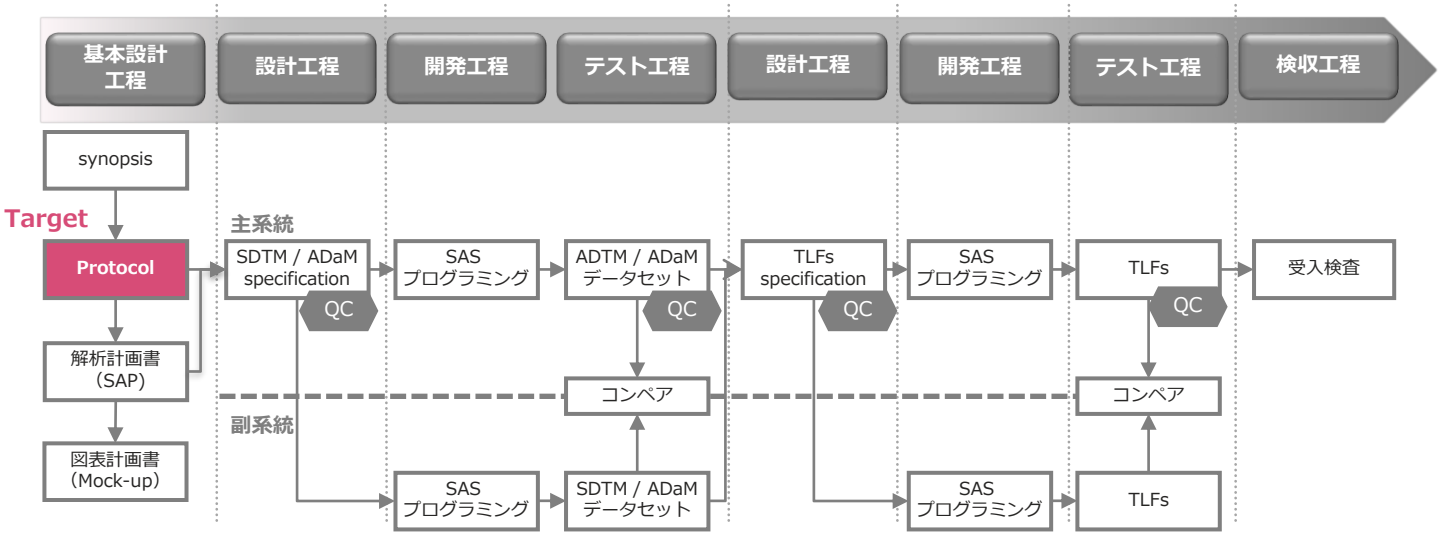
- LLMs always accurate?
  LLMs generate text probabilistically, which means they may produce both factual and incorrect information. This phenomenon is known as hallucination.

- LLMs do not self-learn
  Large language models do not become smarter simply by being used more.
  LLMs like traditional machine learning models, have separate processes for training and inference.

- LLMs and Copyright
  Discussions about copyright issues in handling LLMs are currently ongoing. Various perspectives are being explored regarding the use of copyrighted works as datasets for constructing LLMs and whether the generated results infringe on copyright.

Case Overview

- Streamlining the Creation of Clinical Trial Documents
  - Aim to improve the efficiency of creating documents used in clinical trials.
  - Documents are created based on information from past trials and reference materials.



Preparation

- Dataset for Implementation and Evaluation
  - Past Exam Information
    - Development Phase
    - Domain
  - Related Documents
    - Protocol Template
    - Investigator’s Brochure (IB)
    - Proposal Document
  - References
    - Research Paper
    - Guidelines

Implementation

Types of Generation Approaches

Approach	Example (ex. primary endpoints)

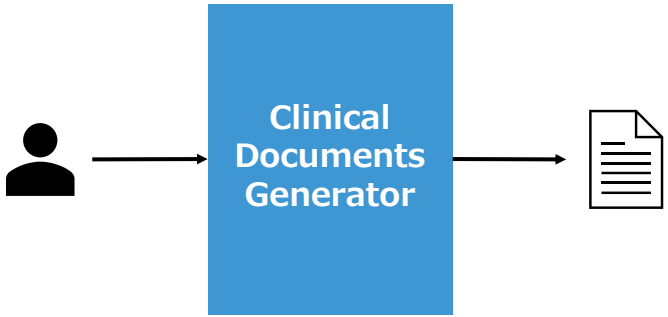
Rule-based Approach	def primary endpoints(target_desease):
	  if target_desease=“Solid Tumor”:
	  Objective Response Rate
	
	  elseif target_desease=“Schizophrenia”:
	  PANSS

Machine Learning Approach	【Training Data】 No.1 Background:”Solid Tumor is ~”, “Objective Response Rate” No.2 Background:”Schizophrenia is ~”, “PANSS” … efficacy_assessments=trained_model({Background:” Vascular endothelial growth factor (VEGF) is a ”})
LLMs Approach	【Reference Document】 ・New response evaluation criteria in solid tumours:Revised RECIST guideline efficacy_assessments=LLMModel({prompt:“What are the primary endpoints for solid tumors?”})

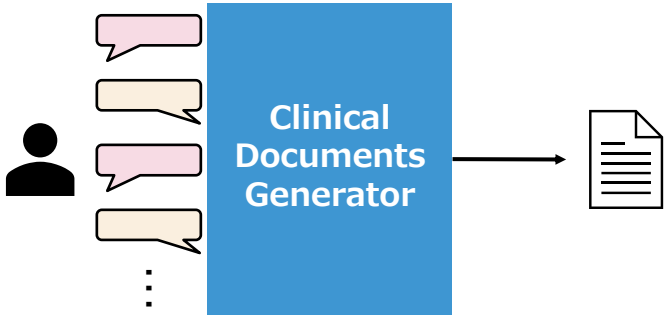
Types of Execution Approaches

Two execution methods were prepared based on the characteristics of the chapter to be generated.

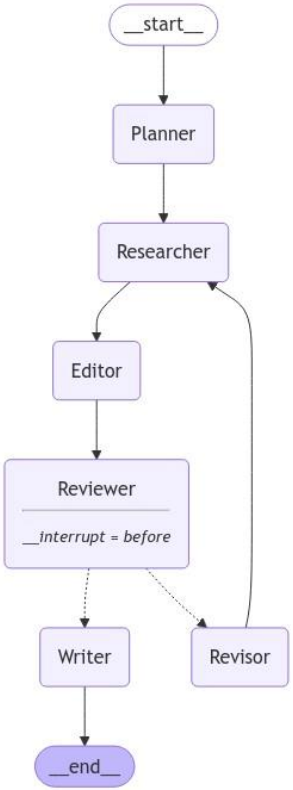
Batch Execution



Interactive Execution



Implementation



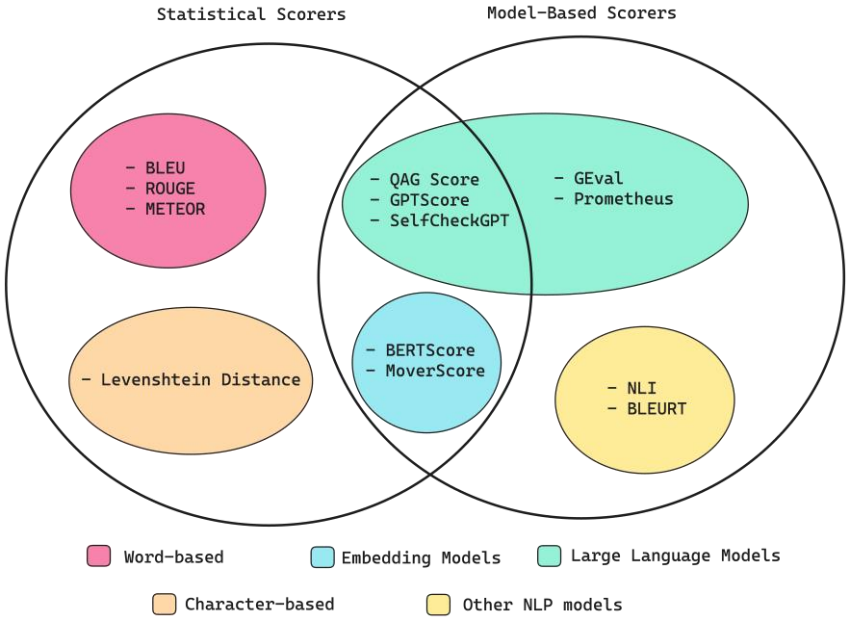
- Interactive Execution : **Multi-Agent**

  Interactively generate text using multi-agent systems and user feedback.

  - **Planner**: Generates the initial outline
  - **Researcher**: Conducts research based on the initial outline (RAG)
  - **Editor**: Edits the outline based on the research results
  - **Reviewer**: Reviews the outline through human evaluation
  - **Revisor**: Revises the outline based on the review feedback
  - **Writer**: Writes the main text

Quantitative Evaluation

Quantitative evaluation is effective for recursively evaluating created texts. However, as there are no absolute metrics for text evaluation, it is recommended to use quantitative evaluation as one of the evaluation criteria.



https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Quantitative Evaluation

BLEU Score

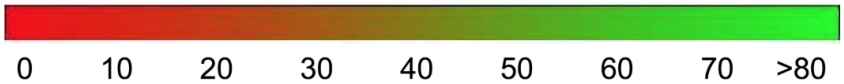
The BLEU score is a commonly used metric for evaluating machine translation models in natural language processing. Since it can be calculated automatically, it is applicable to large datasets. However, it may fail to fully capture meaning and fluency, as matching words and phrases alone may not adequately evaluate the quality of the text.

Interpretation

As a rough guideline, the following interpretation of BLEU scores (expressed as percentages rather than decimals) might be helpful.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

The following color gradient can be used as a general scale interpretation of the BLEU score:



https://cloud.google.com/translate/docs/advanced/automl-evaluate

Qualitative Evaluation

Quantitatively evaluating the quality of text generated by LLMs is challenging, so human evaluation is often conducted. Human evaluators assess aspects such as the quality, fluency, relevance, and creativity of the generated text.

The diversity of quantitative evaluation studies

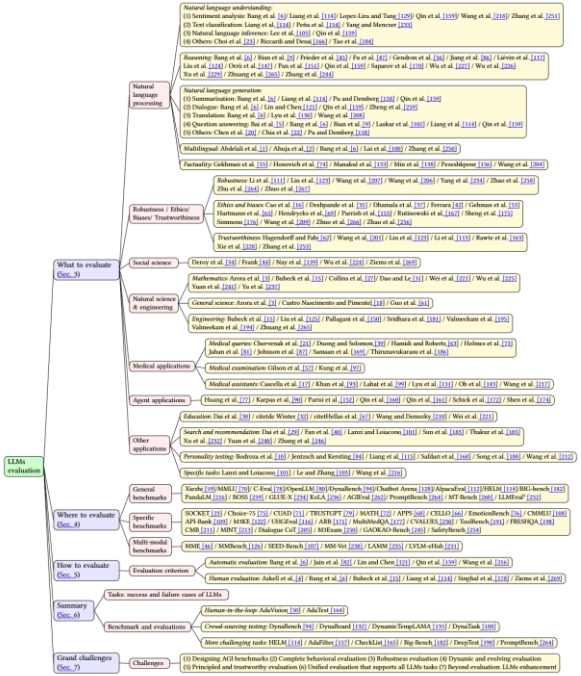


Fig. 1. Structure of this paper.

https://arxiv.org/pdf/2307.03109

Qualitative Evaluation

For this task, as a significant portion involves writing based on references and past exam information, qualitative evaluation was conducted by referring to evaluation methods for summarization tasks. The evaluation was carried out using established criteria for accuracy and fluency.

Accuracy

Addition Unnecessary and irrelevant snippets from the source are included in the summary.

Omission Key point is missing from the output.

Inaccuracy Intrinsic Terms or concepts from the source are misrepresented and thus unfaithful.

Inaccuracy Extrinsic The summary has content not presented in the source and factually incorrect.

Positive-Negative Aspect The output summary represents positive statements whereas the source segment is negative, and vice versa.

Fluency

Duplication A word or longer portion of the text is repeated unnecessarily.

Word Form Problems in the form of a word, including agreement, POS, tense-mood-aspect, etc.

Word Order Problems in the order of syntactic constituents of a sentence.

https://aclanthology.org/2020.emnlp-main.33/

Output Example

Background Section

以下が試験計画書の背景章の本文です：

乳癌は世界中で女性が罹患する最も一般的ながんであり、その中でもホルモン受容体陽性HER2陰性早期乳癌は最も一般的なタイプである。乳癌患者の90%以上が早期疾患で診断され、そのうち約70%がホルモン受容体陽性HER2陰性の乳癌である。このタイプの乳癌の特性は、早期および遅延の遠隔再発を予測することができるため重要である。再発後の内分泌応答性は再発後の生存改善のための重要な要素であり、ER陽性、HER2陰性の再発乳癌において転移腫瘍が内分泌耐性であるかどうかを確認することが重要である。

現在の乳癌治療には限界があり、新たな取り組みが必要である。早期乳癌の多くの患者が化学療法で過剰治療されていることが明らかになっており、これが補助化学療法の意思決定を支援するための遺伝子発現シグネチャーの開発につながった。しかし、これらの方法もまた限界があり、新たな取り組みとさらなる研究が必要である。

既存の研究では、多遺伝子アッセイ（MammaPrint/BluePrint）がホルモン受容体陽性HER2陰性早期乳癌の治療に有用であることが示されている。MammaPrintは、臨床的に高リスクと評価された患者でも化学療法なしで優れた結果を示す低リスク患者を特定する能力を示し、補助化学療法の意思決定におけるテストの価値を確認した。また、BluePrintは80の遺伝子の発現に基づいて活性化した下流の分子経路を捉え、腫瘍を機能的なサブタイプに分類する。

本研究の目的は、ホルモン受容体陽性HER2陰性早期乳癌の早期再発リスクのより精緻な評価を可能にすることである。早期再発リスクの高い患者を早期に特定し、適切な治療を施すことで、患者の生存率を向上させ、社会的負担を軽減することを目指している。本研究の新規性は、既存の研究結果に基づいて、より精緻なリスク評価を可能にすることにある。これにより、ホルモン受容体陽性HER2陰性早期乳癌の新たな治療法の道を開く可能性がある。

※発表用のデモ文章であり、実案件の内容ではありません

Results

- In five months, 46% (70/151 subsections) of the protocol document implementation has been completed.
  - The PoC period is included.
  - Each chapter has been adapted to implement rule-based, machine learning, and LLM approaches accordingly.
- It has been confirmed that the effort required for document generation has been reduced by 50% compared to when performed manually.
- Challenges remain in the full automation of all chapters of the protocol.
  - There are still areas highlighted in qualitative evaluations that require attention.
  - There are still chapters that have not been implemented.

Future work

- Generated text is never perfect
  - It was found that while generated text can be brought closer to the correct answer, achieving perfection is difficult. Increasing the number of chapters that can accommodate generation logic also incurs additional costs.
  - Implementation improvements can enable the acquisition of missing information for generation. However, subtle nuances and expressions that may be evaluated differently depending on the person are continuously subject to detailed critiques.
  - Balancing the trade-off between cost and quality is not straightforward.

- Change the Utilization Method
  - Documents previously created by humans have been used as reference data to generate texts closer to the correct answers. An alternative approach could involve establishing a workflow where the generated document is utilized as a draft, with humans refining and editing it as necessary.

Summary

- Developed a system for creating protocol documents using LLM.
- Conducted the construction and evaluation process for the LLM application.
- The generated results demonstrated a quality suitable for use in production operations.
- Expanding quality and scope involves cost challenges. There is room to explore usage methods, including technical solutions or improvements to business workflows.

THANK YOU

Make society's wishes come true through IT.

TIS INTEC
Group