



University of Stuttgart
Institute of Industrial Automation
and Software Engineering

Development and evaluation of a knowledge management system powered by Large Language Models

Jiayi He

Electrical Engineering

Supervisor: Yuchen Xia M.Sc.



- **Motivation and Problems**
- **Basis**
- **Conceptual Design**
- **Implementation**
- **Evaluation of LLMs**
 - **Dataset Creation**
 - **Results and Findings**
- **Summary and Outlook**

Motivations

🔍 Navigating through massive volumes of previous research.

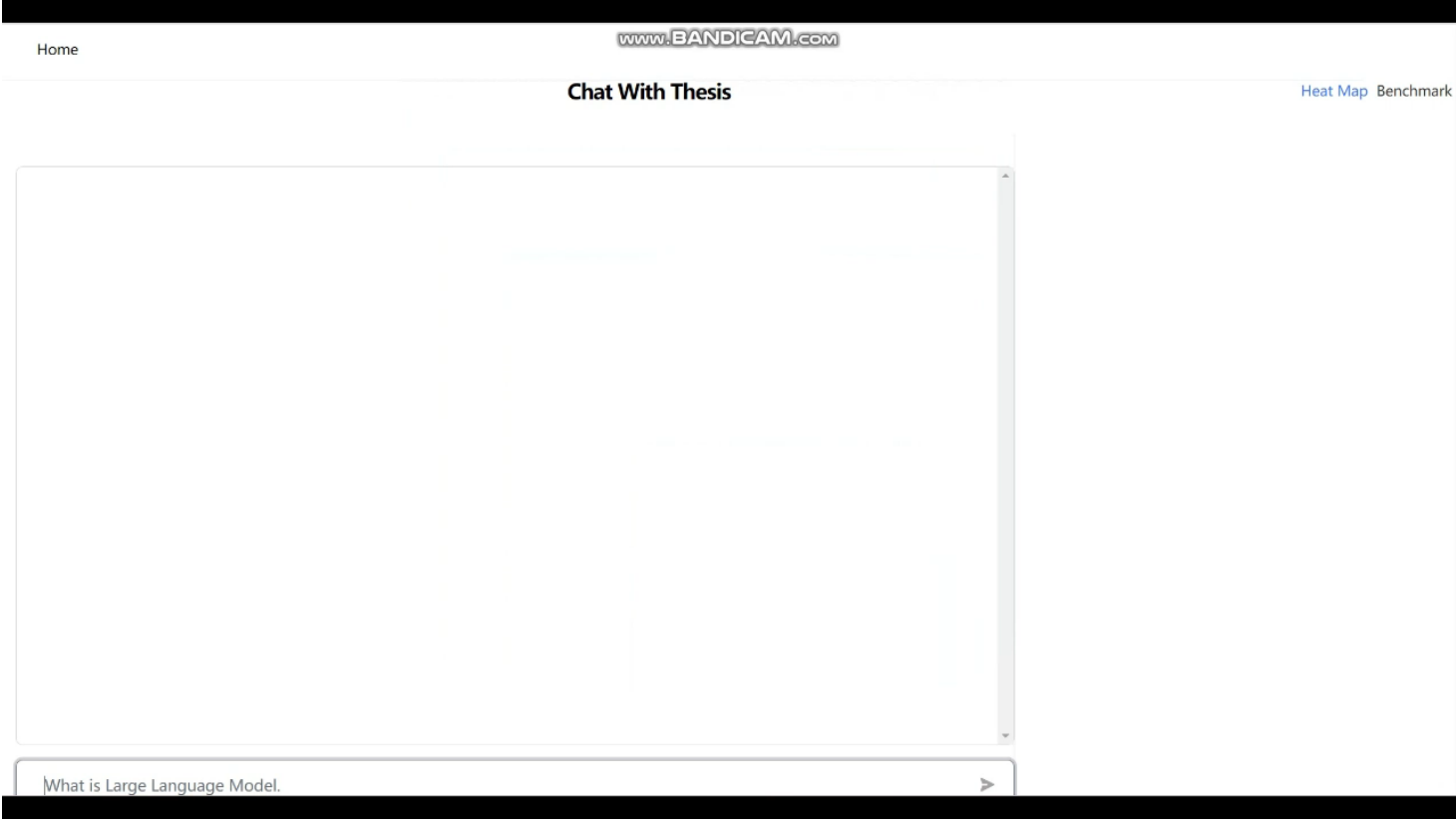


[1]

KMS Benefits:

- **Efficiency:** Streamline reuse of IAS researches.
- **Ease:** Minimize supervisor workload.
- **Collaboration:** Better synergy in industrial automation domain.

KMS Demo Video



Two Project Focuses

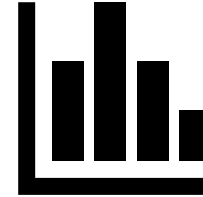
- Design & **Implementation**:
Knowledge **M**anagement **S**ystem development



KMS integrated with LLM



- **Scientific** Research Focus:
LLM performance evaluation



Benchmark for evaluating LLMs

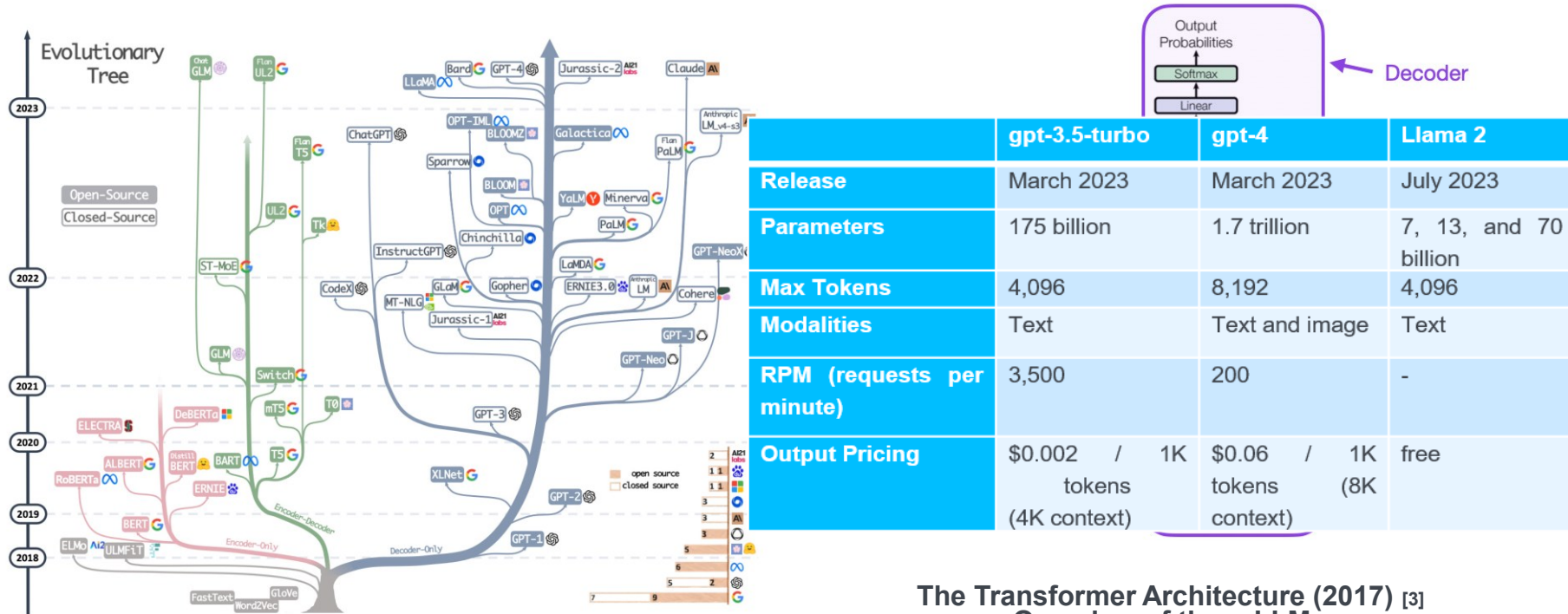
- **Motivation and Problems**
- **Basis**
- **Conceptual Design**
- **Implementation**
- **Evaluation of LLMs**
 - **Dataset Creation**
 - **Results and Findings**
- **Summary and Outlook**

Design focus

Scientific focus

Basis

Large Language Models



The Transformer Architecture (2017) [3]
 Overview of three LLMs
 • Attention Mechanism

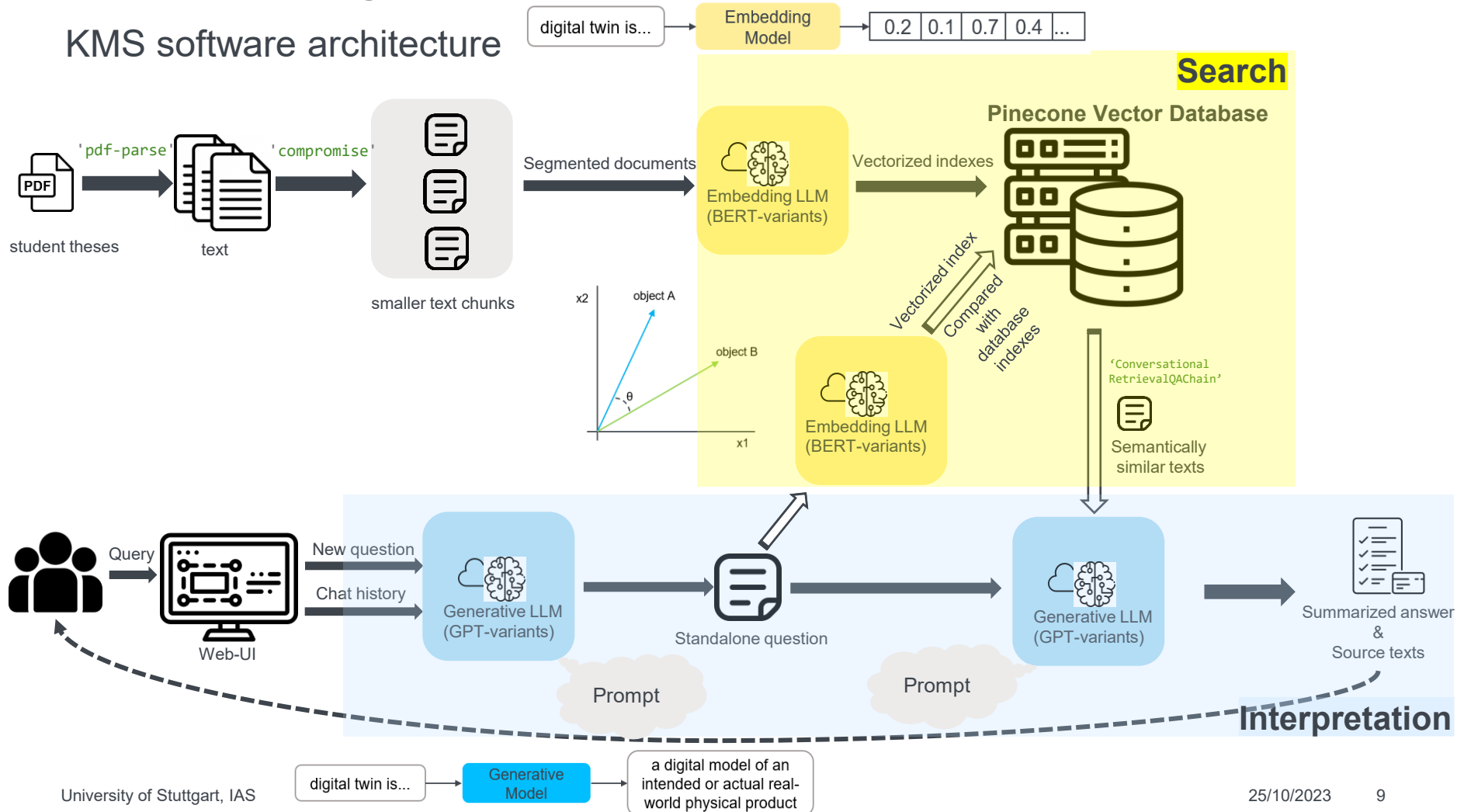
- **Motivation and Problems**
- **Basis**
- **Conceptual Design**
- **Implementation**
- **Evaluation of LLMs**
 - **Dataset Creation**
 - **Results and Findings**
- **Summary and Outlook**

Design focus

Scientific focus

Conceptual Design

KMS software architecture



- **Motivation and Problems**
- **Basis**
- **Conceptual Design**
- **Implementation**
- **Evaluation of LLMs**
 - **Dataset Creation**
 - **Results and Findings**
- **Summary and Outlook**

Design focus


Scientific focus


Implementation Result

User Interface

Home

Chat With Thesis

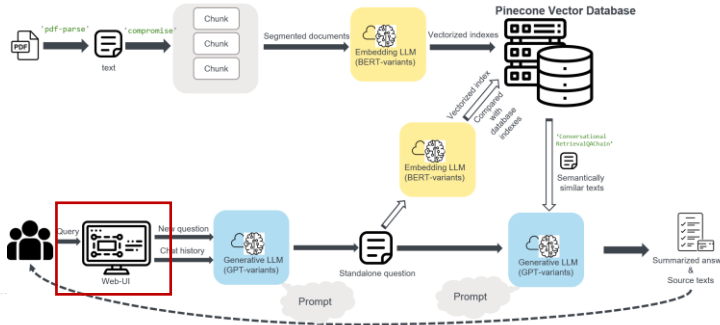
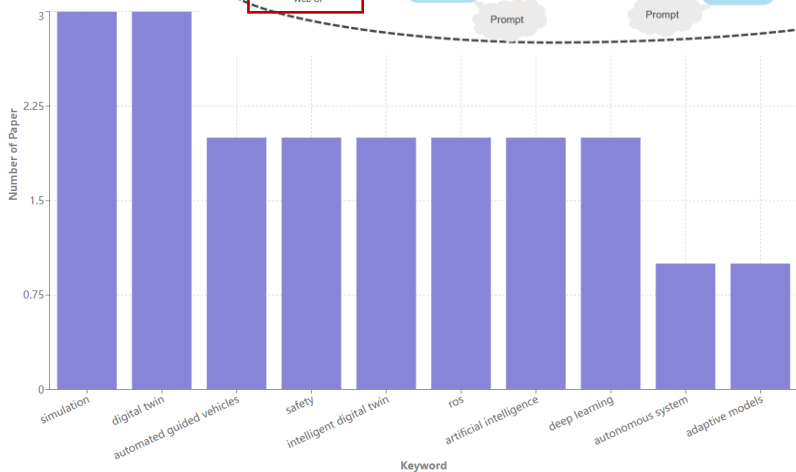
 What is the main purpose of using a Digital Twin?

 The main purpose of using a Digital Twin is to create a virtual representation of a physical entity (such as a product, process, o and optimize its behavior and performance. This allows for better decision-making, improved product development, and mon

What is Large Language Model.

Chat Area

User Input Area



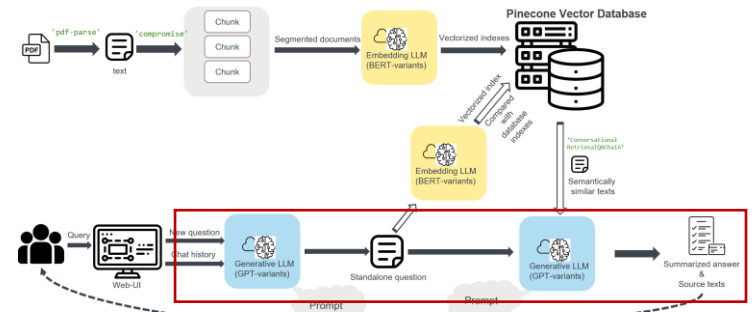
Source 5

Source 6

Reference Section

Implementation

Interpretation by LLM



What is a digital twin?

A digital twin is an executable virtual model of physical equipment that can be continuously updated throughout its lifecycle. It represents the properties, condition, and behavior of the physical object through models and data. It provides real-time information for decision-making and can make predictions about the future behavior of the physical object. The digital twin acts as a bridge between the virtual world and the real physical world, allowing analysis of physical entities and enabling applications such as virtual commissioning, fault prediction, and reconfiguration planning.

- Response generated by KMS learning from the database.

How can I make lasagna?

The ability to reject unknown question

I'm sorry, but I am tuned to only answer questions related to the provided context.

- Response generated by KMS when no relevant information was found in the database.

Implementation

Similarity Search

References

Source 1

Source 2

Digital Twin to have the same behaviour and functionality as the physical object.

Intelligent System Any formal or informal system to manage data gathering, to obtain and process data, to interpret data, and to provide reasoned judgements to decision makers as a basis for action. ISO 23247 Standard of a Digital Twin framework for manufacturing.

Ontology A set of concepts and categories in a subject area or domain that shows their properties and relations between them.

Prototype The first example of something, such as a machine programme, from which all later forms are developed.

Sensor Device which detects or measures a physical or chemical property and records, indicates, or otherwise responds to it.

VDI 2206 Standard for the development of mechatronic and Cyber-Physical Systems.

VDI 3682 Standard for formalised process descriptions.

xiv Abstract The concept of the Digital Twin is increasingly used in manufacturing industry to support

From: Development and Application of a Context Model for the Digital Twin of an Intelligent Warehouse by **Christian Alexander Köhler Elektrotechnik und Informationstechnik**

Source 3

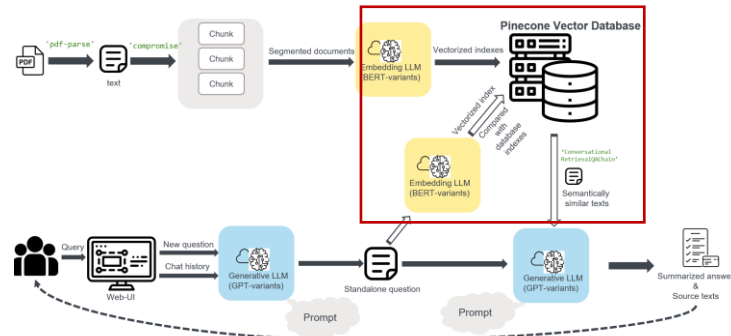
Source 4

Source 5

Source 6

Similarity Search Result

Reference to the Source Documents



ients the properties, condition, and behavior of the physical object through models and data. In acts as a bridge between the virtual world and the real physical world, allowing analysis

the entire lifecycle of the equipment. When the term was first introduced, the technology was associated with large datasets. The manufacturing and process technology industries have a history of using data-driven technologies that include reliable sensors, high-speed networks, low-cost data storage, and Big Data analytics. The digitization of the digital twin. With the current digitalization trend and demand for customized manufacturing style [6]. A digital twin can assist industrial manufacturers in reducing

system to manage data gathering, to obtain and process data, to interpret data, and to provide of concepts and categories in a subject area or domain that shows their properties and relationships which detects or measures a physical or chemical property and records, indicates, or otherwise describes. xiv Abstract The concept of the Digital Twin is increasingly used in manufacturing

```

question What is a digital twin?
{
  text: Retrieved text virtual
  It provides a decision-m
  of physical entities and choosing applications su
  sourceDocuments: [
    Document {
      pageContent: 'A digital twin is an ex
      was not accepted instantly across the indus
      digital twin concepts without mentioning them expl
      Data analytics. In addition, key technologies re
      mized, high-quality products in highly variable b
      costs, improving customer service, and identifyin
      metadata: [Object],
      relevanceScore: 9
    },
    Document {
      pageContent: 'Digital Twin to have the same
      e reasoned judgements to decision makers as a bas
      ions between them. Prototype The first example of
      is responds to it. VDI 2286 Standard for the dev
      g industry to support',
      metadata: [Object],
      relevanceScore: 8
    },
    Document {

```

- **Motivation and Problems**
- **Basis**
- **Conceptual Design**
- **Implementation**
- **Evaluation of LLMs**
 - **Dataset Creation**
 - **Results and Findings**
- **Summary and Outlook**

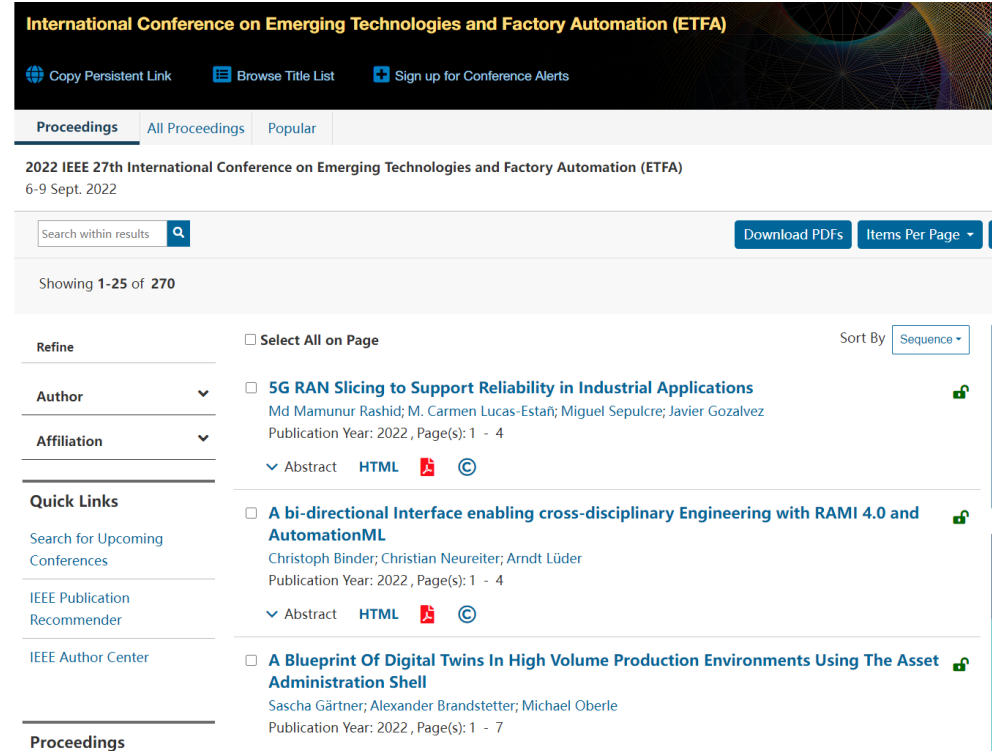
Design focus

Scientific focus

Evaluation Dataset Creation

Generating Questions Related To Industrial Automation

- Evaluation Dataset
 - Derived from ETFA 2022 concepts.
 - Result: Over 900 questions across industrial automation fields.



The screenshot displays the IEEE ETFA 2022 conference website. At the top, the conference title "International Conference on Emerging Technologies and Factory Automation (ETFA)" is shown in yellow on a dark background. Below the title are links for "Copy Persistent Link", "Browse Title List", and "Sign up for Conference Alerts". A navigation bar includes "Proceedings", "All Proceedings", and "Popular". The main content area shows the conference details: "2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)" and "6-9 Sept. 2022". A search bar with the text "Search within results" and a magnifying glass icon is present, along with buttons for "Download PDFs" and "Items Per Page". Below the search bar, it says "Showing 1-25 of 270". On the left, there is a "Refine" section with expandable categories for "Author" and "Affiliation". Below this is a "Quick Links" section with links to "Search for Upcoming Conferences", "IEEE Publication Recommender", and "IEEE Author Center". The main list of proceedings includes three entries, each with a checkbox, title, authors, publication year, page range, and a lock icon. The first entry is "5G RAN Slicing to Support Reliability in Industrial Applications" by Md Mamunur Rashid, M. Carmen Lucas-Estañ, Miguel Sepulcre, and Javier Gozalvez. The second entry is "A bi-directional Interface enabling cross-disciplinary Engineering with RAMI 4.0 and AutomationML" by Christoph Binder, Christian Neureiter, and Arndt Lüder. The third entry is "A Blueprint Of Digital Twins In High Volume Production Environments Using The Asset Administration Shell" by Sascha Gärtner, Alexander Brandstetter, and Michael Oberle. Each entry has a dropdown menu for "Abstract" and links for "HTML", a PDF icon, and a Creative Commons icon.

International Conference on Emerging Technologies and Factory Automation (ETFA)

Copy Persistent Link Browse Title List Sign up for Conference Alerts

Proceedings All Proceedings Popular

2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)
6-9 Sept. 2022

Search within results

Download PDFs Items Per Page

Showing 1-25 of 270

Refine

Select All on Page

Sort By Sequence

Author

Affiliation

Quick Links

Search for Upcoming Conferences

IEEE Publication Recommender

IEEE Author Center

Proceedings

5G RAN Slicing to Support Reliability in Industrial Applications
Md Mamunur Rashid; M. Carmen Lucas-Estañ; Miguel Sepulcre; Javier Gozalvez
Publication Year: 2022, Page(s): 1 - 4

Abstract HTML PDF CC

A bi-directional Interface enabling cross-disciplinary Engineering with RAMI 4.0 and AutomationML
Christoph Binder; Christian Neureiter; Arndt Lüder
Publication Year: 2022, Page(s): 1 - 4

Abstract HTML PDF CC

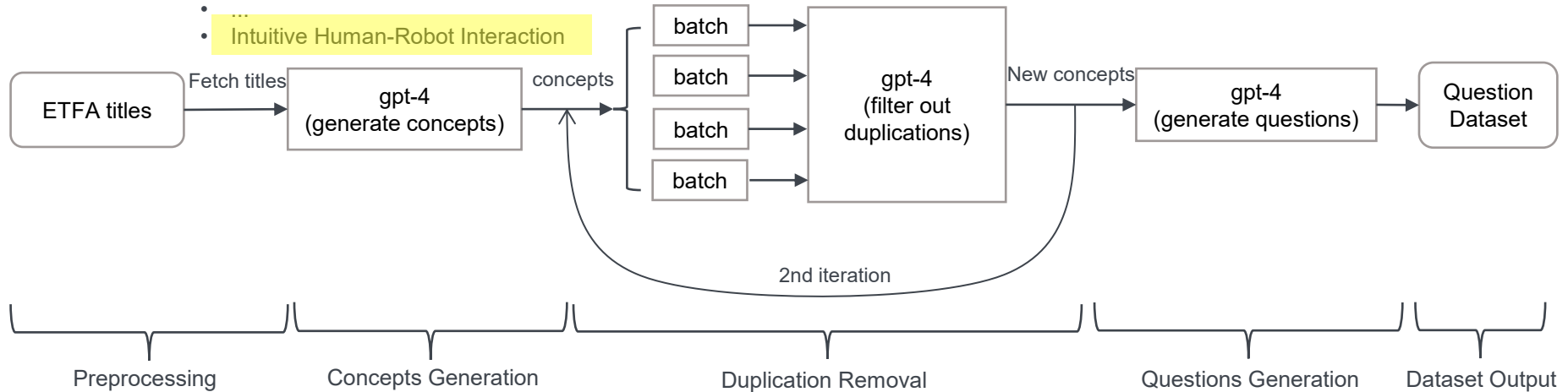
A Blueprint Of Digital Twins In High Volume Production Environments Using The Asset Administration Shell
Sascha Gärtner; Alexander Brandstetter; Michael Oberle
Publication Year: 2022, Page(s): 1 - 7

Benchmark Development for Assessing Different LLMs

Evaluation Dataset Creation

A framework for safe and intuitive human-robot interaction for assistant robotics

- Intuitive Human-Robot Interaction
- Human-Robot Interaction Framework
- Assistant Robotics Framework
- Safe and Intuitive Interaction for Robotics
- ...
- Intuitive Human-Robot Interaction



976
Questions

- How can human-robot interaction be made more intuitive?
- What is the Human-Robot Interaction Framework?
- What is the purpose of the Assistant Robotics Framework?
- How can robotics be designed to ensure safe and intuitive interaction with humans?

Evaluation of the Large Language Models

Evaluation Criteria

- Quantitative metrics (Grading with 0-5)

Solid ideas
survive criticism



How to calculate
these metrics?

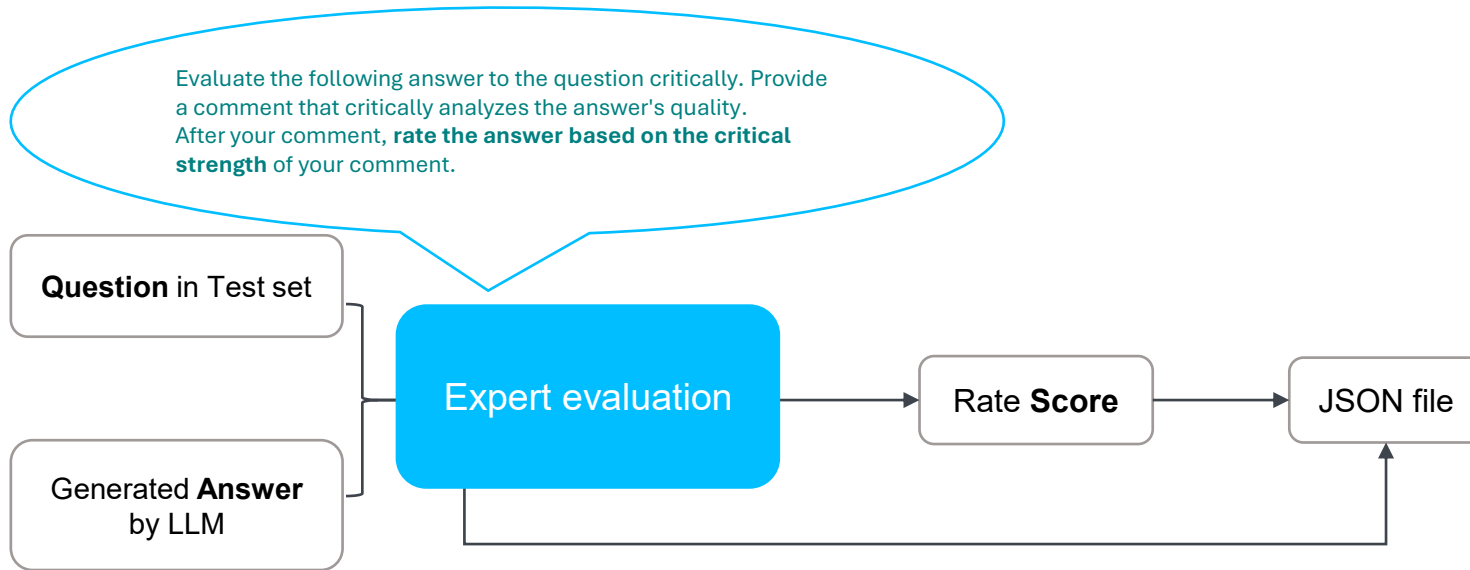
- The more reasonable the criticism is, the poorer the answer is.

Evaluation of the Large Language Models

gpt-4 as the Judge

<<Judging LLM-as-a-judge with MT-Bench and Chatbot Arena>> (Jul. 2023):

“...GPT-4 can achieve an agreement rate of over 80%,
on par with the level of agreement among human experts.”



```
"question": "44.\tWhat is AutomationML?",  
"answer": "AutomationML (Autoregressive Machine Learning) is a type of machine learning algorithm that is specifically designed for time series forecasting tasks. It is a type of neural network.",  
"rate": 1.0,  
"comment": "The answer provided seems well-structured and detailed, but it is misleading and incorrect.  
The answer actually describes a form of Machine Learning algorithm incorrectly labeled as AutomationML when in fact AutomationML stands for Automation Markup Language, not Autoregressive Machine Learning. AutomationML is an open initiative to create an open, non-proprietary, XML-based data format as an interface for plant engineering tools. It aims to simplify the exchange of plant design and control information between different engineering tools used in plant design and operation. It combines and integrates different establish
```

Evaluation of the Large Language Models

Results

100 samples
1500 evaluations

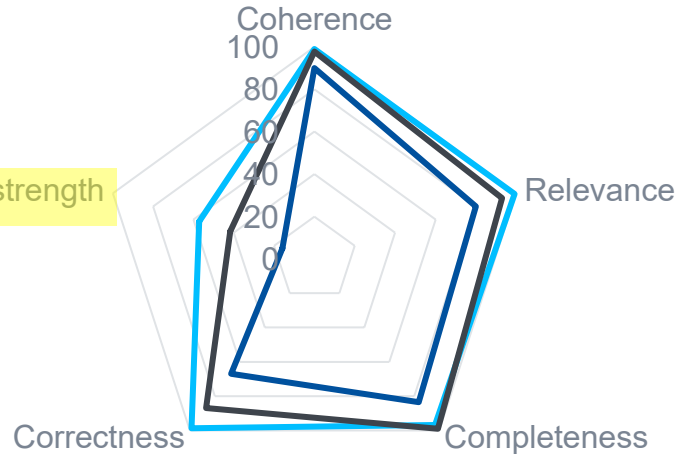
Generative models	Coherence	Relevance	Completeness	Correctness	Critical comment strength
gpt-3.5-turbo	99.0	99.2	96.8	98.7	57.2
LlaMA-2-7B	90.0	80.0	83.5	66.9	15.9
LlaMA-2-70B	97.8	93.0	99.0	87.0	41.8
GPT4 (reference)	100	100	100	100	100

LLM response evaluation results

Results Visualization

— gpt-3.5-turbo — LLaMA-2-7B — LLaMA-2-70B

Critical comment strength



- gpt-3.5 has a better performance than Llama-2 in most dimensions.
- Llama-2-7B has a relatively bad performance due to its limited training data.
- gpt-4 is capable of 'judging', and providing critical insights.

- **Motivation and Problems**
- **Basis**
- **Conceptual Design**
- **Implementation**
- **Evaluation of LLMs**
 - **Dataset Creation**
 - **Results and Findings**
- **Summary and Outlook**

Design focus

Scientific focus

Summary and Outlook

Summary:

- A KMS is developed
 - Easier information retrieval.
- Evaluation Benchmark
 - Synthetic questions.
 - Evaluation metrics.

Limitation:

- Only 31 theses in the KMS database.

Outlook:

- Benchmark used for evaluating new LLMs in the future.



University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!

Jiayi He

e-mail st176731@stud.uni-stuttgart.de

phone +49 (0) 711 685-

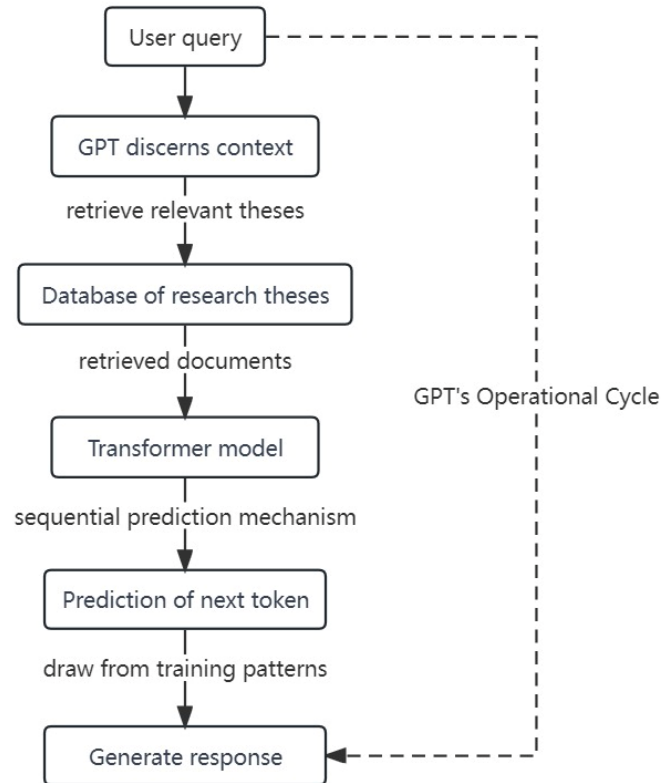
fax +49 (0) 711 685-

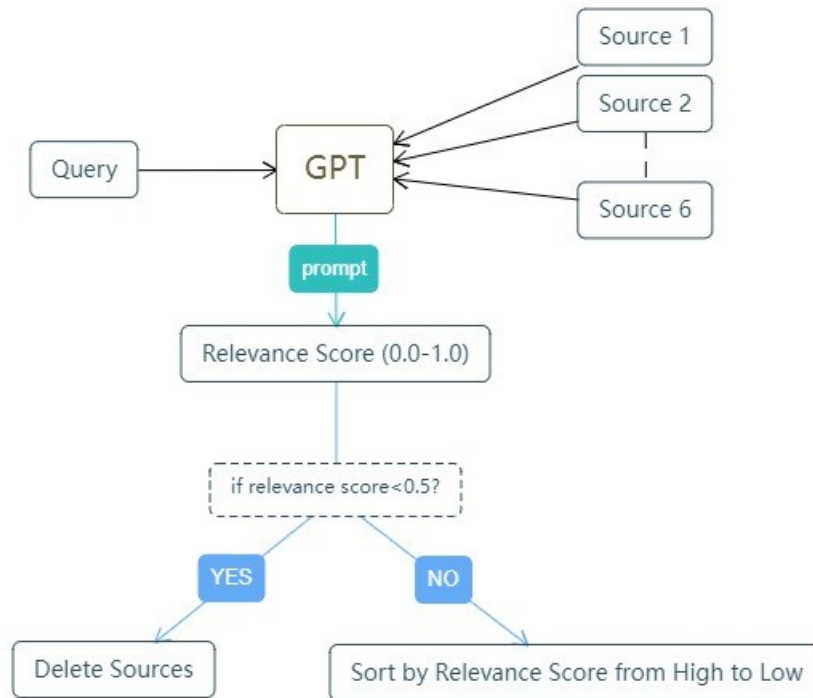
University of Stuttgart



Quelle

1. <https://twimlai.com/podcast/twimlai/topic/large-language-models-llms/>.
2. <A Survey of Large Language Models>, arXiv: 2303.18223v1.
3. <Fine-tuning of general neural language models for automated interpretation of the semantics of the data properties in industrial automation domain>, Master Thesis, Zhewen Xiao.
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, “Attention Is All You Need,” Comput. Res. Rep., pp. 6000–6010, 2017.





LLM as Knowledge Manager

Two Scenarios

Chat with LLM

- A vast ocean of generalized knowledge.
- Lack of help in newest specific knowledge domain.

```
"id": 13,  
"question": "\tHow can human-robot interaction be made more intuitive?",  
"answer": "There are several ways in which human-robot interaction can be made more intuitive:  
\n\n1. Natural Language Processing: Integrating robust natural language processing capabilities in robots can enable them to understand and respond to human commands.  
\n\n2. Gestures and Expressions: Incorporating sensors and cameras that can detect human gestures and facial expressions can enable robots to interpret non-verbal communication.  
\n\n3. Simplified Interfaces: Designing user interfaces with simplicity in mind helps in making the interaction more intuitive. Avoiding complex menus and excessive information.  
\n\n4. Visual Feedback: Providing visual feedback to users helps in creating a better understanding of the robot's actions and intentions. Displaying information clearly and concisely.  
\n\n5. Context Awareness: Robots that can understand and adapt to the context in which they are operating can provide more intuitive interactions. By perceiving the user's environment and needs.  
\n\n6. Personalization: Allowing users to personalize the robot's behavior and appearance can enhance the intuitive connection between humans and robots. Customizing the robot's responses based on user preferences.  
\n\n7. Training and Learning: Implementing machine learning algorithms can enable robots to learn from user interactions and adapt their behavior accordingly. This allows the robot to become more familiar with the user's preferences and needs.  
\n\nOverall, combining advances in technology with user-centered design principles can significantly enhance the intuitive nature of human-robot interaction. The goal is to create a seamless and natural interaction experience for both humans and robots."  
"rate": 5.0,  
"comment": "The answer is coherent and well-organized. The writer strategically structures the different methods to make human-robot interaction more intuitive and easy to understand."
```

Chat with KMS

- Database contains research theses at IAS.
- Information retrieval in specific domain.

```
"id": 13,  
"question": "\tHow can human-robot interaction be made more intuitive?\r",  
"answer": "Based on the given context, it is not directly mentioned what can be done to make human-robot interaction more intuitive."  
"rate": 1.0,  
"comment": "The given answer is not coherent with the question asked. Instead of providing any suggestions or insights on how human-robot interaction can be made more intuitive, the answer simply states that it is not directly mentioned in the given context.""
```