University of Stuttgart
Institute of Industrial Automation
and Software Engineering

Survey on Large Language Models for Applications in Industrial Automation and Software Engineering

Research thesis 3655

Haonan Chen

Electrical Engineering and
Information Technology

Contents

- **State-of-the-art** of LLMs
- **Methodological approaches**
- **Applications**
- **Evaluation methods**
- **Challenges and future directions**
- **Conclusion**

State-of-the-art LLM

What the latest LLMs have been released?

- Search Strategy
- Filtering Strategy
- Overview

State-of-the-art of LLMs

Search Strategy

1. Review of Existing Surveys and Literature

	Related Survey Papers
1	A Survey of Large Language Models
2	A Survey on Large Language Models for Software Engineering
3	A Survey on Large Language Models: Applications,Challenges, Limitations, and Practical Usage
4	Large Language Models: A Survey
…	…
13	Large Language Models for Software Engineering: A Systematic Literature Review

2. Keyword Searches

LLMs + Text Generation / Text2Text Generation / Multimodal

3. Industry news

OpenAI, MISTRAL AI_, HUAWEI, Microsoft, Google AI

4. AI model Databases and Aggregators

Papers With Code

Hugging Face

State-of-the-art of LLMs

Filtering Strategy

LLMs released from **January 1, 2023 to May 31, 2024**

↓

LLMs with **significant impact within the AI community**

↓

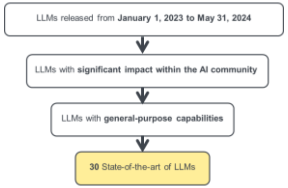
LLMs with **general-purpose capabilities**

↓

30 State-of-the-art of LLMs

State-of-the-art of LLMs

Overview



- **30** State-of-the-art of LLMs are collected

Model	Institute	Release_Time	Size(B)	Pre-training Data Scale	Open Source
Llama	Meta	Feb-2023	7-65	1.4T	YES
Kosmos-1	Microsoft	Feb-2023	1.6B	360B	
GPT-4	OpenAI	Mar-2023			NO
PanGu-Σ	Huawei	Mar-2023	1085	329B	NO
PaLM-E	Google	Mar-2023	12,84,562		NO
Stanford Alpaca	Stanford University	Mar-2023	7	52k instruction-following demonstrations	YES
Vicuna	LMSYS	Mar-2023	13	70k instruction-following demonstrations	YES
WizardLM	Microsoft,Peking University	Apr-2023	7B	250K	YES
PaLM2	Google	May-2023			NO
Phi-1	Microsoft	Jun-2023	1.3	7B	YES
Llama 2	Meta	Jul-2023	Llama 2-7B: 7 B Llama 2-13B: 13 B Llama 2-34B: 34 B (not released) Llama 2-70B: 70 B	2T	YES
Baichuan2	Baichuan	Sep-2023	Baichuan 2-7B:7B Baichuan 2-13B:13B	2.6T	YES
QWEN	Alibaba	Sep-2023	Qwen 1.5-1.8B:1.8B Qwen 1.5-7B:7B Qwen 1.5-14B:14B	3T	YES
FLM	Beijing Academy of Artificial Intelligence	Sep-2023	FLM-16B:16B FLM-51B:51B FLM-101B:101B	0.31T	YES
Phi-1.5	Microsoft	Sep-2023	1.3	30B	YES
LongChat	UC Berkeley, University of Washington, UC San Diego, CMU, MBZUAI, and USC	Oct-2023	13		YES
Skywork	Kunlun	Oct-2023	13	3.2T	YES
Mistral 7B	Mistral	Oct-2023	7		YES
Falcon	Technology Innovation Institute	Nov-2023	Falcon-7B:7B Falcon-40B:40B Falcon-180B:180B	3.5T	YES
Mixtral 8x7B	Mistral	Dec-2023	47		YES
Phi-2	Microsoft	Dec-2023	2.7	1.4T	YES
Gemini	Google	Dec-2023	Gemini Nano-1:1,8B Gemini Nano-2:3.25B		NO
Qwen 1.5	Alibaba	Feb-2024	0.5B,1.8B,4B,7B,14B,32B,72B ,110B		YES
Jamba	AI21 Labs	Mar-2024	52		YES
Mixtral 8x22B	Mistral	Apr-2024	141		YES
Llama 3	Meta	Apr-2024	8,70	15T	YES
Phi-3	Microsoft	Apr-2024	Phi-3 Mini:3,8B Phi-3 Medium:14B	3.3T	YES
OpenELM	Apple	Apr-2024	1.1B	1.8T	YES
FILM7B	Xi'an Jiaotong University,Microsoft,Peking University	Apr-2024	7B		YES
WizardLM2	Microsoft	Apr-2024	7B,70B,8x22B		YES

7B? 13B?

65B?

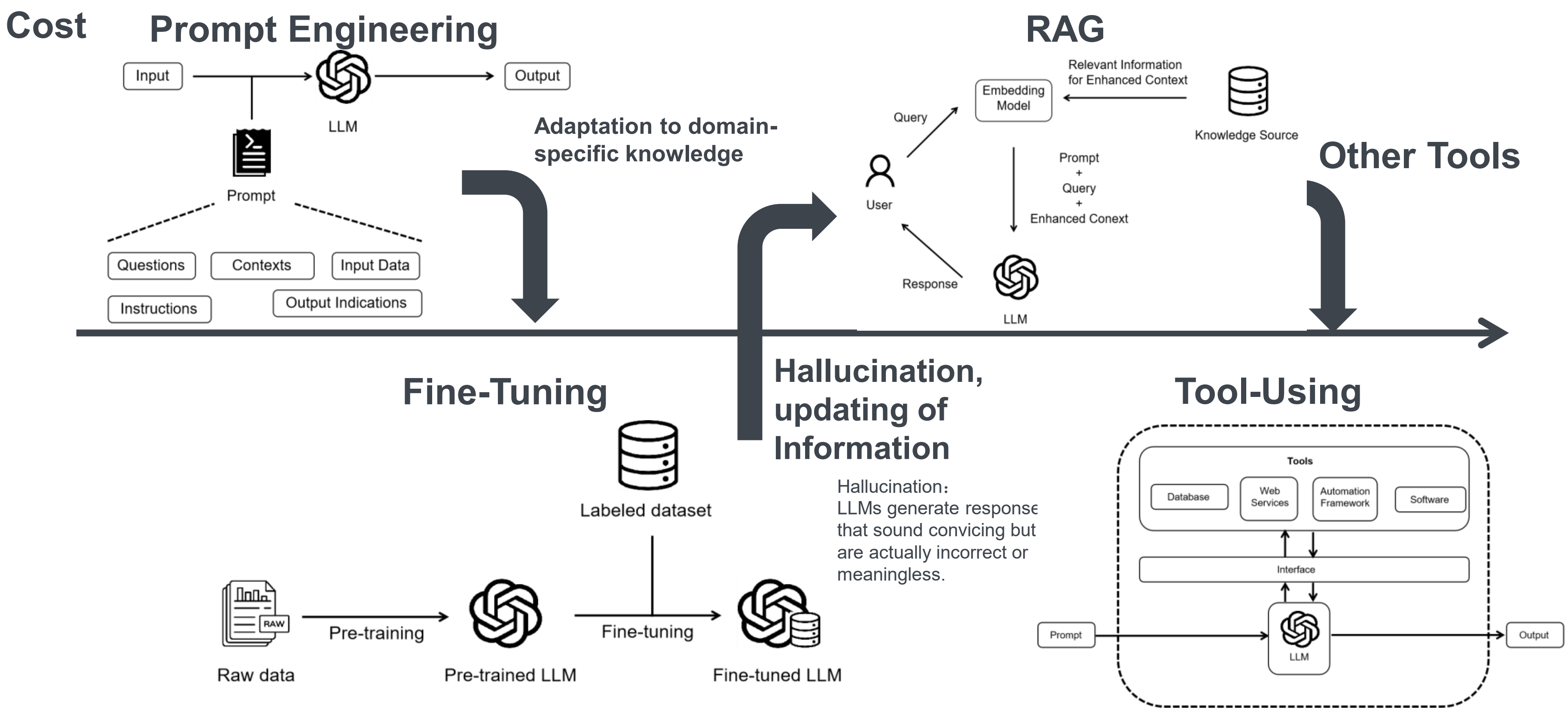
Methodologies

But, how can we apply these LLMs?

- Prompt Engineering
- Fine-Tuning
- RAG
- Tool-Using

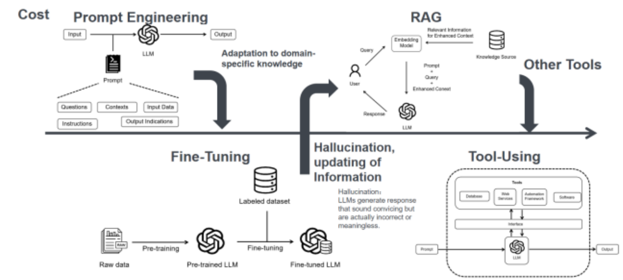
Methodological approaches using LLMs

Prompt Engineering, Fine-Tuning, RAG, and Tool-Using



Methodological approaches using LLMs

Prompt Engineering vs Fine-Tuning vs RAG



	Promt Engineering	Fine-Tuning	RAG
Cost	Low	High	Medium
Convenience of Use	High	Low	Medium
Model Customisation Requirement	Low	High	Medium
Hallucination	Low	Medium	High
Frequency of Updates	Low	Variable	High

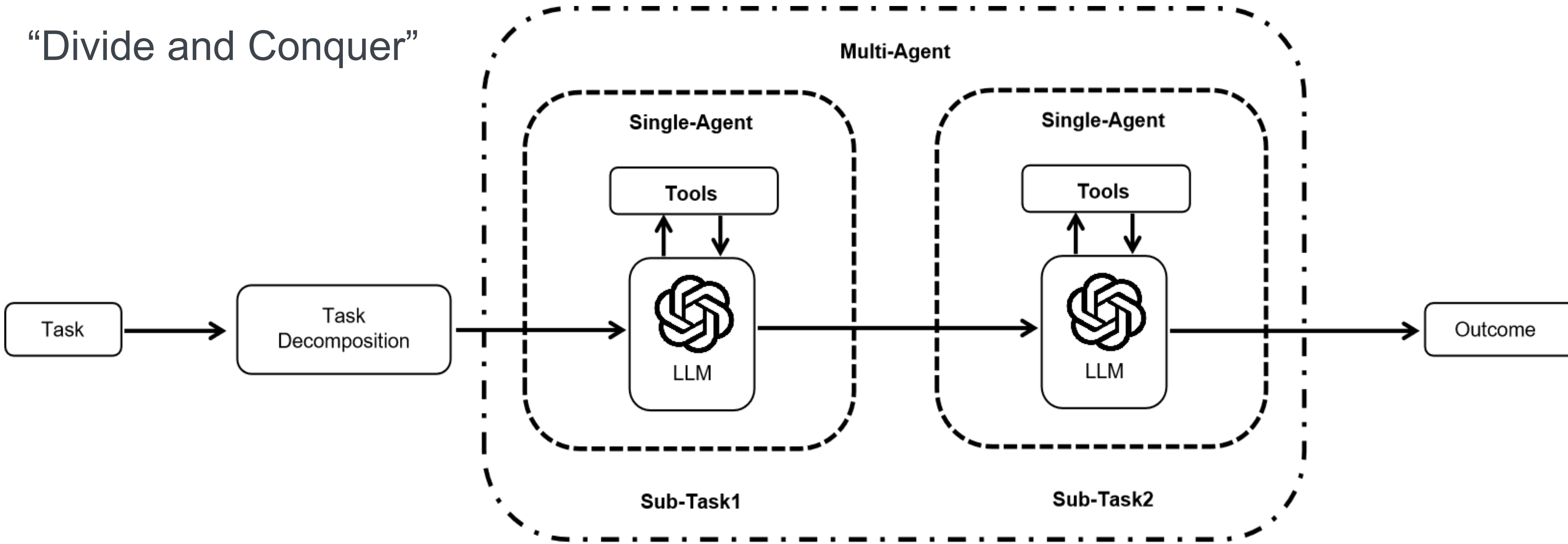
Advanced Methodologies

Agent design

Methodological approaches using LLMs

LLM Agents

“Divide and Conquer”



Applications

What downstream tasks are LLMs spread to?

- 109 application cases in Software Engineering
- 68 application cases in Industrial Automation

Applications of LLMs in industrial automation and software engineering

Applications in Software engineering

Classification Criteria

During the classification process, **ChatGPT-4** is utilized to evaluate **the relevance of each application case to these categories**.

Each case was scored on a scale from 0 to 10.

If an application case had **scores of 8 or higher in other categories**, those categories were also noted.

This approach ensures that while each case is categorized into its primary functional area, its relevance to other areas is also acknowledged.

Applications of LLMs in industrial automation and software engineering

Applications in Software engineering

Overview

Categ	tion cases
Code (	
Softwa	
Softwa Autom	
AI-Driv	
Simula	
Comm	

(a) The Protoss strategically transitions to a Void Ray-centric composition, adept at countering the might of Zerg's Ultralisks.

Reference:

- Ma W, Mi Q, Yan X, et al. Large language models play starcraft ii: Benchmarks and a chain of summarization approach[J]. arXiv preprint arXiv:2312.11865, 2023.
- Qian C, Cong X, Yang C, et al. Communicative agents for software development[J]. arXiv preprint arXiv:2307.07924, 2023.

Applications in Industrial Automation

- Classification based on function
- Classification based on role

Applications of LLMs in industrial automation and software engineering

Applications in Industrial Automation Engineering

Two different classification methods

- **Function-based Classification**

This method categorizes the applications based on the specific functions that LLMs perform within industrial automation, highlighting their distinct contributions.

- **Role-based Classification**

This approach analyzes each application case to determine which real-world jobs can be supplemented or replaced by LLMs, illustrating their practical impact on workforce dynamics.

Applications … gineering

Appl… on
Over…

Cate…
Auto…
Infor…
Hum… Assis…
Auto…
Data…
Dom…
Intell…

r

Reference:

- Wang L, Ling Y, Yuan Z, et al. Gensim: Generating robotic simulation tasks via large language models[J]. arXiv preprint arXiv:2310.01361, 2023.
- Wu J, Antonova R, Kan A, et al. Tidybot: Personalized robot assistance with large language models[J]. Autonomous Robots, 2023, 47(8): 1087-1102.

Role-based Classification

Which real-world job can be supplemented or replaced by LLMs?

Applications of LLMs in industrial automation and software engineering

How LLM can assist human engineer? Or even replace some jobs?

Overview

Role	Number of relevant application cases
Automation Engineer	59
Control Systems Engineer	59
Robotic Engineer	30
Test Engineer	13
Systems Integration Specialist	10
Simulation Engineer	7
Documentation Specialist	6
Maintenance Technician	5
Data Analyst	5
Systems Modeling Engineer	3
R&D Engineer	1

Evaluation

How are LLMs evaluated in SE and IAE?

- 27 Metrics
- 110 Benchmarks

Evaluation of LLMs and LLM-powered Systems

Metrics

Table 21: Robustness and Fairness Metrics

	Metrics	Key Characteristics
1	HEQ[251]	Equivalence to human performance
2	Expected calibration error[252]	Quality of model calibration
3	MCC[253]	Quality of classification
4	AUC [253]	Discriminatory power in binary classification
5	ASR(Attack Success Rate)[254]	Adversarial robustness of LLMs
6	IoU[251]	Overlap between predicted and actual spans

Table 22: Efficiency and Performance Metrics

	Metrics	Key Characteristics
1	Pass@k[75]	Success rate within top K attempts
2	AvgPassRatio[99]	Average success rate across multiple attempts
3	ES(Edit Similarity)[79]	Edit distance similarity of text
4	WAV(Weighted Average Value)[190]	Word alignment accuracy
5	Proof success rate	Proof success rate
6	Inference Time	Inference Time
7	Model Size	Model Size

Table

Categ
Task-
Langu
Robu
Effici

1
2
3
4
5
6
7
8
9

C
G
M
M
8
C
C

Accuracy of Understanding, the ability of information retrieval

Machine translation, sentiment analysis, or summarization

Benchmark

- Model-level Benchmarks
- Applications-level Benchmarks

Evaluation of LLMs and LLM-powered Systems

Benchmarks implementation

- The collected benchmarks are categorized into two groups: **Model-level benchmarks** and **applications-level benchmarks**.
- **Model-level benchmarks**: fundamental capabilities of the model

  Language understanding, reasoning
- **Applications-level benchmarks**: the model's performance in tasks that closely mimic real-world applications

  Code generation, understanding informal dialogue
- **56 model-level benchmarks** are collected and **54 applications-level benchmarks** are collected

Evaluation of LLMs and LLM-powered Systems

Benchmarks

Categories	Number of relevant Benchmarks	Representative Benchmarks
Model-level Benchmarks	56	Language understanding (MMLU), Open-ended questions answer (QuAC), Commensense reasoning (HellaSwag)
Applications-level Benchmarks	54	Tools manipulation (ToolBench), Agent's interaction with users (AgentBench), Code generation (HumanEval)

- MMLU - Massive Multitask Language Understanding
- QuAC - Question Answering in Context
- HellaSwag - Commen sense reasoning
- ToolBench - Ability to manipulate tools
- AgentBench - LLM-as-Agent's reasoning and decision-making abilities
- HumanEval - Ability to generate code

Challenges and Future directions

How can LLMs be strengthened and optimzed in SE and IAE?

- Data scarcity
- Deployment
- Multi-modal models

Challenges and Future Directions

- **Data Scarcity**

  Lacking datasets specific to industrial processes, difficulty in effective learning and generalization

- **Deployment**

  Resource constraints, the need for low-latency real-time responses, and complex system integration

- **Multi-modal Models**

  Aligning information between modalities

Conclusion

Conclusion

- **30** prominent general-purpose **LLMs**, **177 use cases** in software engineering and industrial automation, **27 metrics** and **110 benchmarks** used for evaluating LLM capabilities.

- **Significant potential in applications in software engineering and industrial automation**

  Code generation, system control, and automation process

- **Future research**

  Robust benchmarks and tests, data scarcity, model hallucination, and the comlexity of downstream tasks

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!

Haonan Chen

e-mail st175447@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart
Institute of Automation and Software Systems
Pfaffenwaldring 47, 70550 Stuttgart