ISCAS 中国科学院软件研究所学术年会暨重点实验室科技活动周 2025第十届 学术论文

RMGenie: An LLM-Based Agent Framework for Open Source Software README Generation

RMGenie：基于大语言模型的开源软件README生成代理框架

崔星 吴敬征 李志远 罗天悦 凌祥

41st IEEE International Conference on Software Maintenance and Evolution (ICSME 2025)

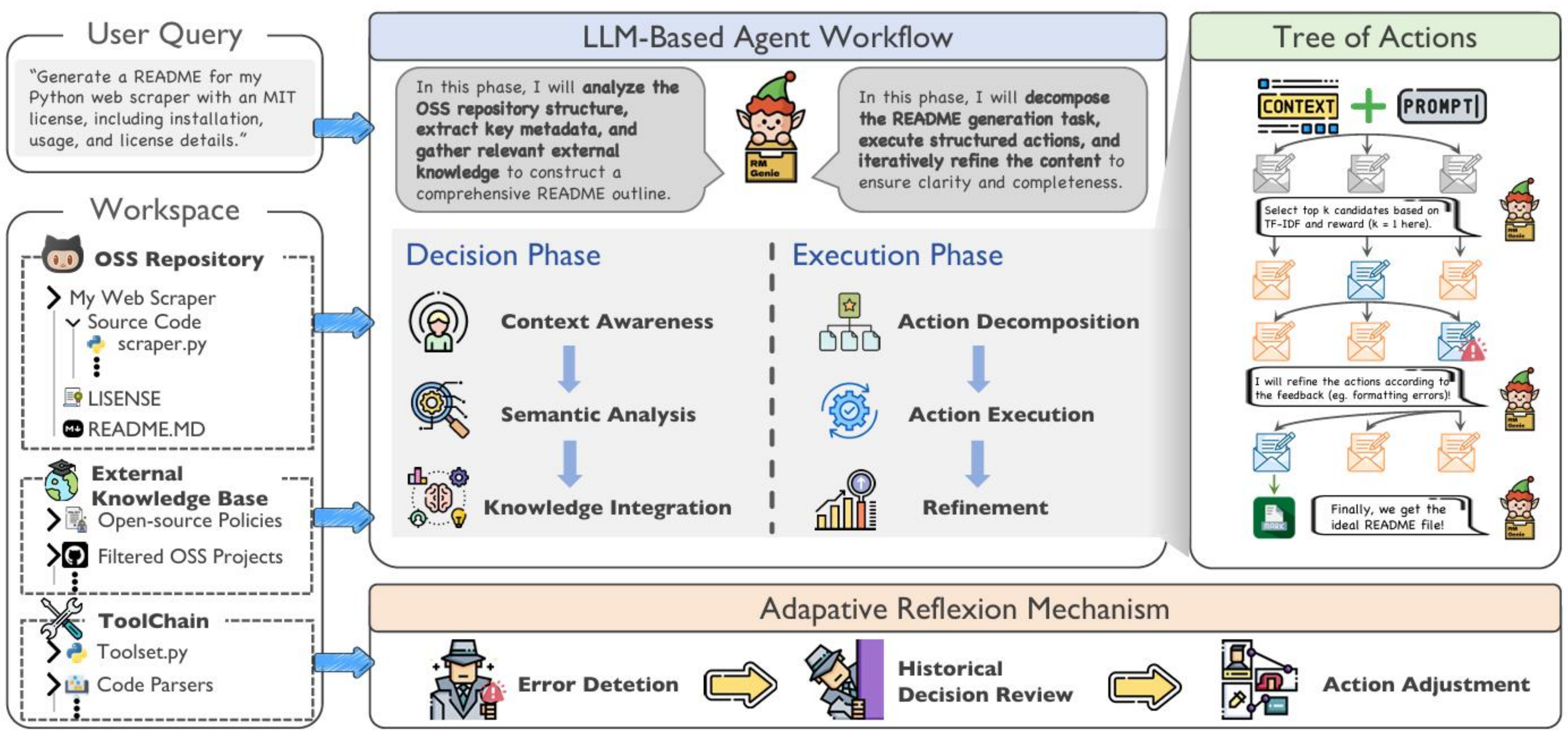
联系人：崔星，13051316652，cuixing@iscas.ac.cn

Background

Open Source Software (OSS) is essential in modern software ecosystems, but around 9.03% of projects lack adequate README documentation, affecting developer efficiency and maintainability. Although recent advancements in natural language processing (NLP) and large language models (LLMs) have automated README generation, challenges remain in integrating real-time knowledge, handling complex software structures, and adapting to project-specific needs. This paper presents RMGenie, an LLM-based agent framework for automated README generation. RMGenie leverages multi-round interactions with external tools to extract key code insights and uses a Tree of Actions model with entropy-based scoring to optimize decision-making. A reflexion mechanism reduces biases and tool errors, while experimental results show RMGenie outperforms existing methods in completeness, instruction adherence, and accuracy.

System Design

RMGenie's design follows a two-phase workflow: the **Decision Phase**, where it analyzes the OSS repository and integrates external knowledge, and the **Execution Phase**, where it decomposes tasks and refines content. A **Tree of Actions (ToA)** model and adaptive reflexion mechanism optimize decision-making and ensure accurate, complete README generation.


User Query
"Generate a README for my Python web scraper with an MIT license, including installation, usage, and license details."
Workspace
OSS Repository
My Web Scraper
Source Code
scraper.py
LISENSE
README.MD
External Knowledge Base
Open-source Policies
Filtered OSS Projects
ToolChain
Toolset.py
Code Parsers
LLM-Based Agent Workflow
In this phase, I will analyze the OSS repository structure, extract key metadata, and gather relevant external knowledge to construct a comprehensive README outline.
RM Genie
In this phase, I will decompose the README generation task, execute structured actions, and iteratively refine the content to ensure clarity and completeness.
Decision Phase
Context Awareness
Semantic Analysis
Knowledge Integration
Execution Phase
Action Decomposition
Action Execution
Refinement
Tree of Actions
CONTEXT
PROMPT
Select top k candidates based on TF-IDF and reward (k = 1 here).
I will refine the actions according to the feedback (eg. formatting errors)!
Finally, we get the ideal README file!
Adapative Reflexion Mechanism
Error Detetion
Historical Decision Review
Action Adjustment

Evaluations

Dimension	Compared LLMs	Win	Tie	Lose	RMGenie Win
Overall Satisfaction	DeepSeek-R1	90	59	1	60.00%
	GPT-4o	91	58	1	60.67%
	ChatGLM4	111	39	0	74.00%
	Qwen2.5	114	36	0	76.00%
	Llama3.1	117	33	0	78.00%
Content Completeness	DeepSeek-R1	119	31	0	79.33%
	GPT-4o	122	28	0	81.33%
	ChatGLM4	124	26	0	82.66%
	Qwen2.5	130	20	0	86.67%
	Llama3.1	132	18	0	88.00%
Factual Accuracy	DeepSeek-R1	99	51	0	66.00%
	GPT-4o	93	55	2	62.00%
	ChatGLM4	108	42	0	72.00%
	Qwen2.5	121	29	0	80.67%
	Llama3.1	115	35	0	76.67%
Instruction Compliance	DeepSeek-R1	77	67	6	51.33%
	GPT-4o	89	60	1	59.33%
	ChatGLM4	106	44	0	70.67%
	Qwen2.5	111	39	0	74.00%
	Llama3.1	113	37	0	75.33%

Table 1: Comparison of RMGenie and LLMs Across Evaluation Dimensions

Model	Score	Overall Satisfaction	Content Completeness	Factual Accuracy	Instruction Compliance
DeepSeek-R1	68.28	68.16	67.5	67.66	69.83
GPT-4o	69.33	69.33	70.33	70.16	67.5
ChatGLM4	67.74	67.5	69.16	66.16	68.16
Qwen2.5	64.41	64.66	65.0	64.66	63.33
Llama3.1	64.03	63.83	64.33	63.33	64.66
RMGenie	**84.33**	**84.83**	**85.33**	**83.66**	**83.5**

Table 2: Comparison on license conflict issues

Method	Evaluation Dimension	Win	Tie	Win Rate
w/o ToA	Overall Satisfaction	86	64	57.33%
	Content Completeness	90	60	60%
	Factual Accuracy	90	60	60%
	Instruction Compliance	84	66	56%

Model	Final Score	Overall Satisfaction	Content Completeness	Factual Accuracy	Instruction Compliance
w/o ToA	80.5	80.5	82.5	77.5	81.5
RMGenie	84.33	84.83	85.33	83.66	83.5

Method	Evaluation Dimension	Win	Tie	Win Rate
w/o Reflexion	Overall Satisfaction	84	66	56%
	Content Completeness	79	71	52.66%
	Factual Correctness	98	52	65.33%
	Instruction Compliance	84	66	56%

Table 3: Ablation study results

Evaluation Dimension	Compared Methods	Win	Tie	Loss	RMGenie Win
Overall Satisfaction	PromptCS	114	36	0	76%
	LARCH	95	49	6	63.33%
	RepoAgent	88	57	5	58.67%
Content Completeness	PromptCS	119	31	0	79.33%
	LARCH	110	40	0	73.33%
	RepoAgent	104	46	0	69.33%
Factual Accuracy	PromptCS	109	41	0	72.67%
	LARCH	99	51	0	66%
	RepoAgent	95	55	0	63.33%

Base Model	Final Score	Overall Satisfaction	Content Completeness	Factual Accuracy	Instruction Compliance
DeepSeek-R1 + ToA + Reflexion	84.33	84.83	85.33	83.66	83.5
GPT-4o + ToA + Reflexion	84.62	84.66	85.16	83.83	84.83
ChatGLM4 + ToA + Reflexion	80.49	80.16	80.5	79.83	81.5
Qwen2.5 + ToA + Reflexion	81.62	81.83	82.00	80.83	81.83
LLama3.1 + ToA + Reflexion	79.12	79.16	78.5	79.66	79.16

Table 4: Automatic and human evaluation of RMGenie

Findings:

- ✓ RMGenie outperforms direct LLM-based methods in key metrics, excelling in content completeness, satisfaction, accuracy, and instruction compliance.
- ✓ Ablations show ToA and reflexion boost RMGenie's accuracy; larger models improve performance but cost more, with GPT-4o and DeepSeek-R1 leading.
- ✓ RMGenie outperforms state-of-the-art methods, achieving the highest scores in satisfaction, completeness, and accuracy with a final score of 84.38.

Contributions

- RMGenie is an LLM-based system for automated README generation, improving documentation adaptability.
- The ToA model enables dynamic reasoning pathways, refining content organization for improved logical coherence and structural clarity beyond static methods.
- RMGenie outperforms existing methods in accuracy and completeness, with top scores in both automated and human evaluations.

Conclusion

RMGenie is an LLM-based framework for automated README generation in OSS, utilizing an LLM agent for analysis and the ToA model for dynamic reasoning. A reflexion mechanism enhances accuracy by refining decisions. Evaluations on 150 GitHub projects show RMGenie outperforms existing tools, and a VSCode plugin integrates it into development workflows, improving efficiency and consistency.

This paper is supported by

Open Source Map