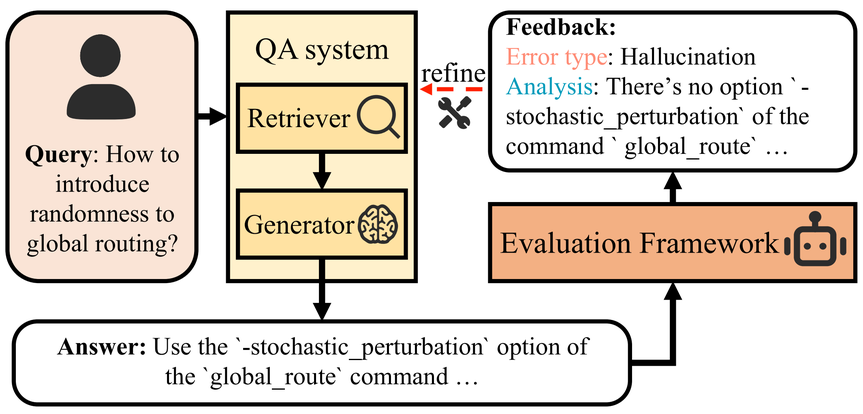
MAEDA: An LLM-Powered Multi-Agent Evaluation Framework for EDA Tool Documentation

Zhenghao Chen[1], Yuan Pu[2,3], Hairuo Han[2], Yuntao Nie[2], Jiajun Qin[2], Yuhan Qin[2], Tairu Qiu[2], Zhuolun He[2,3], Jianwang Zhai[1], Bei Yu[2], Kang Zhao[1]

[1] Beijing University of Posts and Telecommunications, China

[2] The Chinese University of Hong Kong, Hong Kong SAR [3]ChatEDA Tech, China

Introduction



- **EDA documentation is complex and highly interdependent**, making it difficult to navigate and leading to low efficiency and a steep learning curve.
- **Recent advances in LLMs enable EDA QA with RAG**, improving answer reliability by grounding responses in documentation.
- **Evaluation in EDA QA remains limited**, as general-purpose fail to capture domain-specific error types.
- **We propose MAEDA**, a multi-agent framework for fine-grained error identification (retrieval, contradiction, missing, refusal, hallucination), along with a benchmark for systematic evaluation.

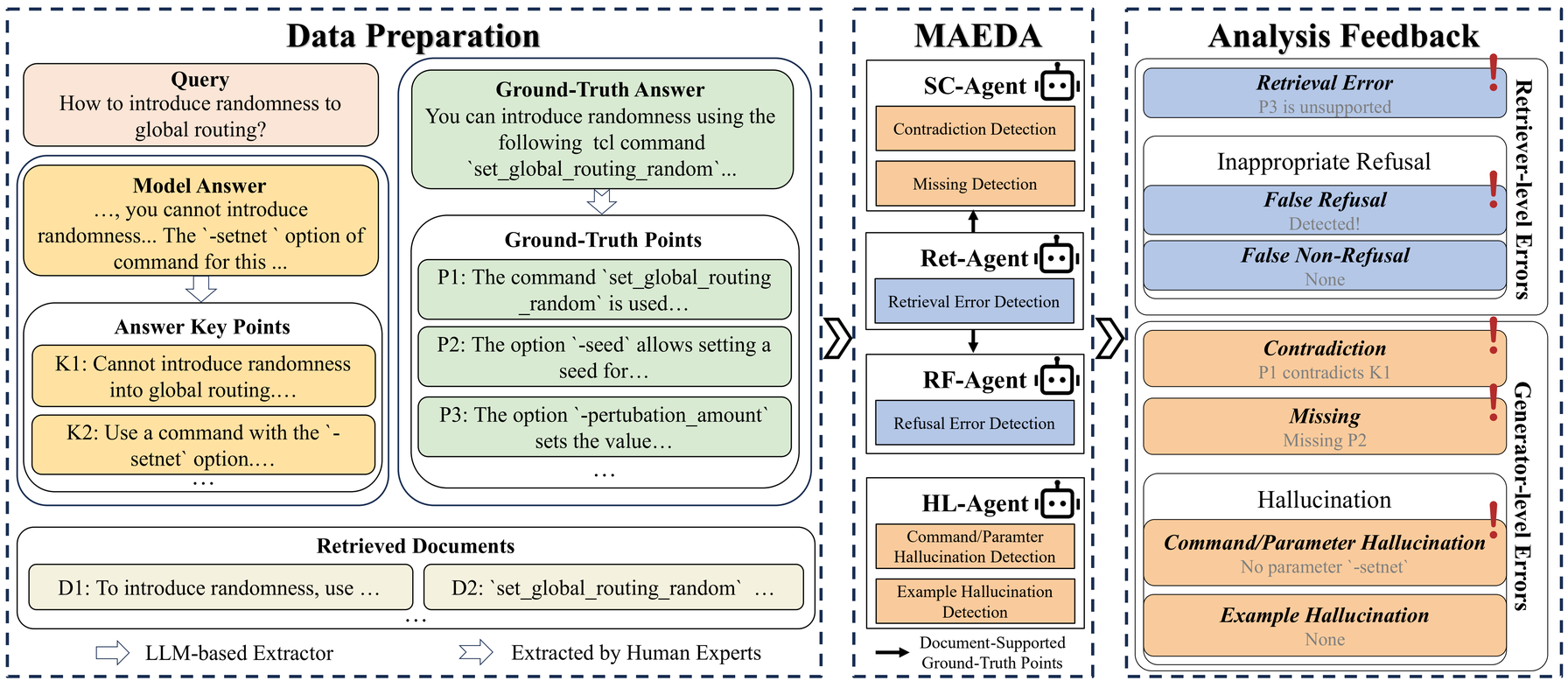
Methodology

● MAEDA

1.MAEDA is a **multi-agent evaluation framework** for RAG-based EDA QA, decomposing evaluation into specialized agents for fine-grained error diagnosis.

2.It represents answers as fine-grained key points, enabling precise **point-to-point alignmen**t with ground truth.

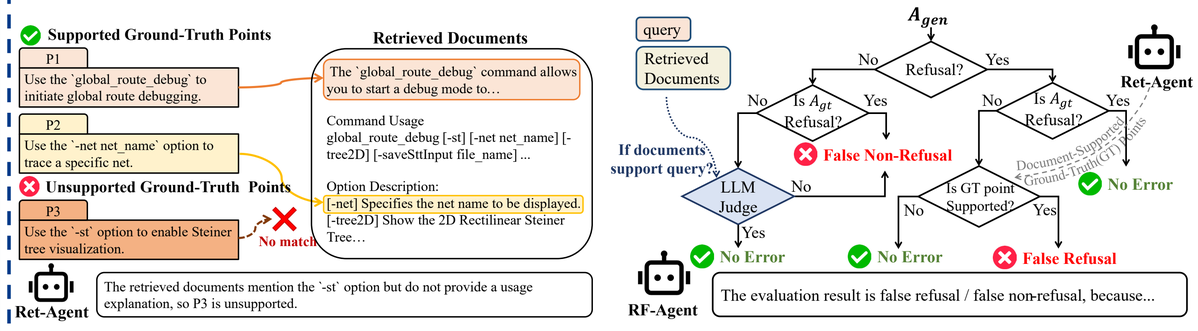
3.The agents **collaborate through structured reasoning** while decoupling error sources, providing interpretable feedback for accurate diagnosis and iterative improvement.



● Benchmark

Existing QA benchmarks **fail to capture fine-grained, domain-specific errors** in EDA scenarios. We introduce a dedicated benchmark based on **OpenROAD**, covering diverse tool functions. It contains **300 QA instances** with positive and negative samples, generated via a **hybrid pipeline** combining real RAG cases and LLM-driven error injection. Each sample is annotated with **expert-defined ground-truth points**, enabling precise and interpretable evaluation.

Retriever-level Errors



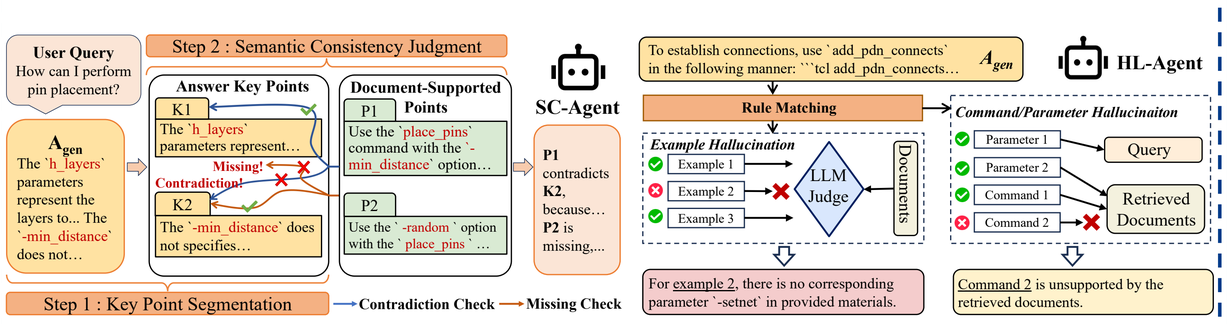
● Retrieval Error Identification Agent

Identifies ground-truth points supported by retrieved documents via point-to-chunk matching. Evaluates coverage and consistency to filter reliable points, providing structured inputs for downstream evaluation.

● Refusal Evaluation Agent

Identifies inappropriate refusal by assessing whether refusal decisions are justified based on retrieved documents. Distinguishes false refusal and false non-refusal via pattern matching and document-grounded reasoning.

Generator-level Errors



● Semantic Consistency Agent

Detects contradictions and missing by aligning answer key points with document-supported ground-truth points through point-to-point matching, using structured reasoning to identify conflicts and insufficient coverage.

● Hallucination Evaluation Agent

Detects hallucinated commands, parameters, and examples via a hybrid evaluation framework, combining deterministic matching and LLM-based structured reasoning under EDA-specific constraints.

Evaluation

TABLE I Model performance on fine-grained error type classification (on OpenROAD Dataset)

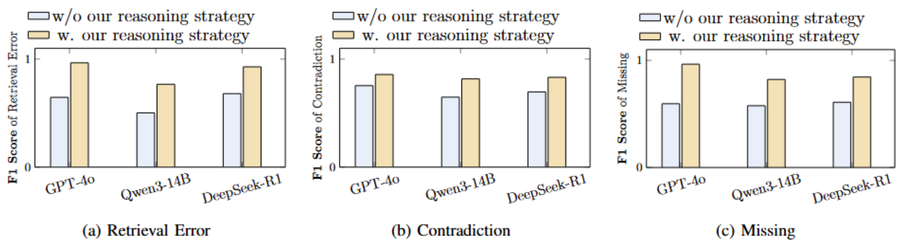
Method	Retrieval Error		Contradiction		Missing		Inappropriate Refusal		Hallucination	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
RAGChecker [34]	0.960	0.692	–		–		–		0.486	0.346
RAGEval [35]	0.958	0.926	0.059	0.039	0.825	0.650	–		0.095	0.132
G-Eval [37]	0.703	0.703	0.881	0.394	0.772	0.550	0.267	0.143	0.385	0.294
MAEDA w. Qwen3-14B	0.747	0.789	0.902	0.676	0.887	0.601	**0.987**	0.711	**0.935**	0.680
MAEDA w. GPT-4o	0.960	**0.947**	0.922	0.758	0.907	0.633	**0.987**	**0.757**	0.860	**0.769**
MAEDA w. Fine-tuned models*	**0.973**	0.936	**0.941**	**0.800**	**0.928**	**0.726**	**0.987**	0.711	**0.935**	0.680

* denotes the use of our fine-tuned models for retrieval error, contradiction and missing, with Qwen3-14B for other error types.

TABLE II Model performance on fine-grained error type classification (on Commercial EDA Tool Dataset)

Method	Retrieval Error		Contradiction		Missing		Inappropriate Refusal		Hallucination	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
RAGChecker [34]	0.821	0.902	–		–		–		0.648	0.246
RAGEval [35]	0.857	0.528	0.033	0.056	0.788	0.788	–		0.038	0.125
G-Eval [37]	0.833	0.641	0.476	0.217	0.852	0.264	0.667	0.08	0.546	0.270
MAEDA w. Qwen3-14B	0.597	**0.976**	0.667	0.625	0.818	0.600	0.827	**0.811**	**0.872**	0.531
MAEDA w. GPT-4o	0.866	0.951	0.767	0.697	0.848	**0.800**	**0.940**	0.758	0.782	**0.550**
MAEDA w. Commercial tool models†	**0.896**	0.952	**0.800**	**0.727**	0.879	0.725	0.827	**0.811**	**0.872**	0.531
MAEDA w. Fine-tuned models*	0.866	0.921	0.700	0.700	**0.909**	0.600	0.827	**0.811**	**0.872**	0.531

† denotes the use of similarly trained commercial EDA tool models for retrieval error, contradiction and missing, with Qwen3-14B for other error types.



(a) Retrieval Error (b) Contradiction (c) Missing

3.Ablation shows that structured CoT reasoning and fine-grained alignment consistently improve performance.

1.MAEDA outperforms baselines across all error types, achieving performance comparable to GPT-4o with advantages in complex reasoning and robustness.

2.Results also validate the effectiveness of hybrid evaluation and demonstrate strong generalization.

Conclusion

We propose MAEDA, a multi-agent evaluation framework for EDA tool documentation QA. By combining domain-specific error types with structured reasoning and fine-grained semantic alignment, it enables accurate and interpretable error identification. We also release an OpenROAD-based benchmark and develop tailored fine-tuning strategies for open-source models. Experiments demonstrate that MAEDA consistently outperforms SOTA methods.

北京郵電大學 Beijing University of Posts and Telecommunications

Contact chenzhenghao@bupt.edu.cn

香港中文大學 The Chinese University of Hong Kong