TechSupportEval: An Automated Evaluation Framework for Technical Support Question Answering

Bohan Chen[1], Yongqian Sun[2], Yuhe Liu[1], Longlong Xu[1], Zhe Xie[1], Changhua Pei[3],
Jing Han[4], Fan Ni[4], Xuhui Cai[5], Ce Yang[5], and Dan Pei[1]

1 2 3 4 5

OUTLINE

Background

Framework

Evaluation

Conclusion

OUTLINE

Background

Framework

Evaluation

Conclusion

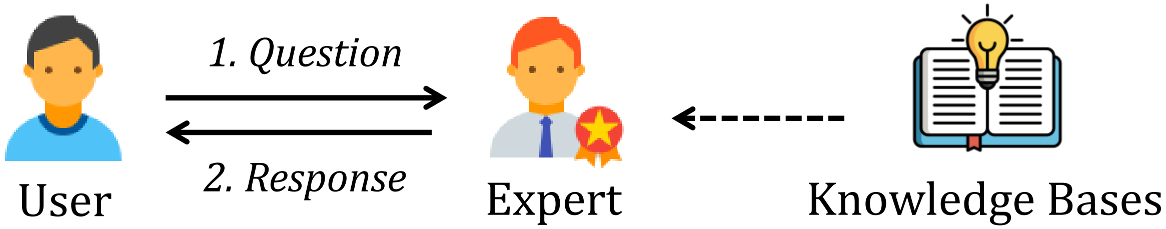
Background

Technical Support Question Answering

Technical Support[1]

Technical support is a service provided to users to diagnose and resolve technical issues to maintain the reliability of IT services.

A common approach: Question Answering (QA)



[1] https://en.wikipedia.org/wiki/Technical_support

Example from Microsoft Forum

Question:

Are there ways to capture the IP addresses that are using my storage accounts in Azure portal?

Response:

Once you have enabled logging for your storage account, the information about operations performed against your storage account is saved in **`$logs` blob container**. It contains a CSV files. The information you're looking for is available in **<requester-ip-address> field**.

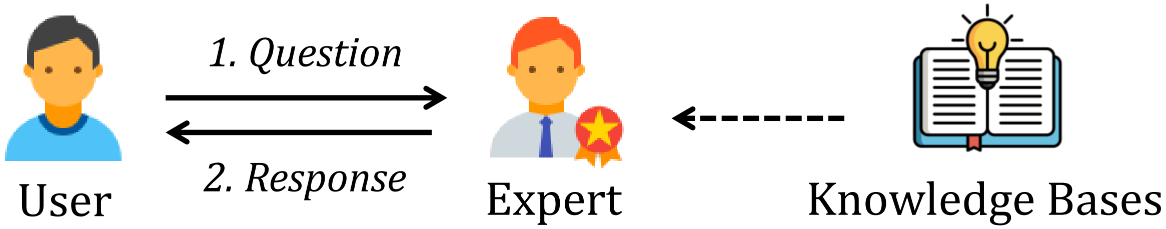
Background

Technical Support Question Answering

Technical Support[1]

Technical support is a service provided to users to diagnose and resolve technical issues to maintain the reliability of IT services.

A common approach: Question Answering (QA)



✓ Accuracy
✗ Latency
✗ Scalability

[1] https://en.wikipedia.org/wiki/Technical_support

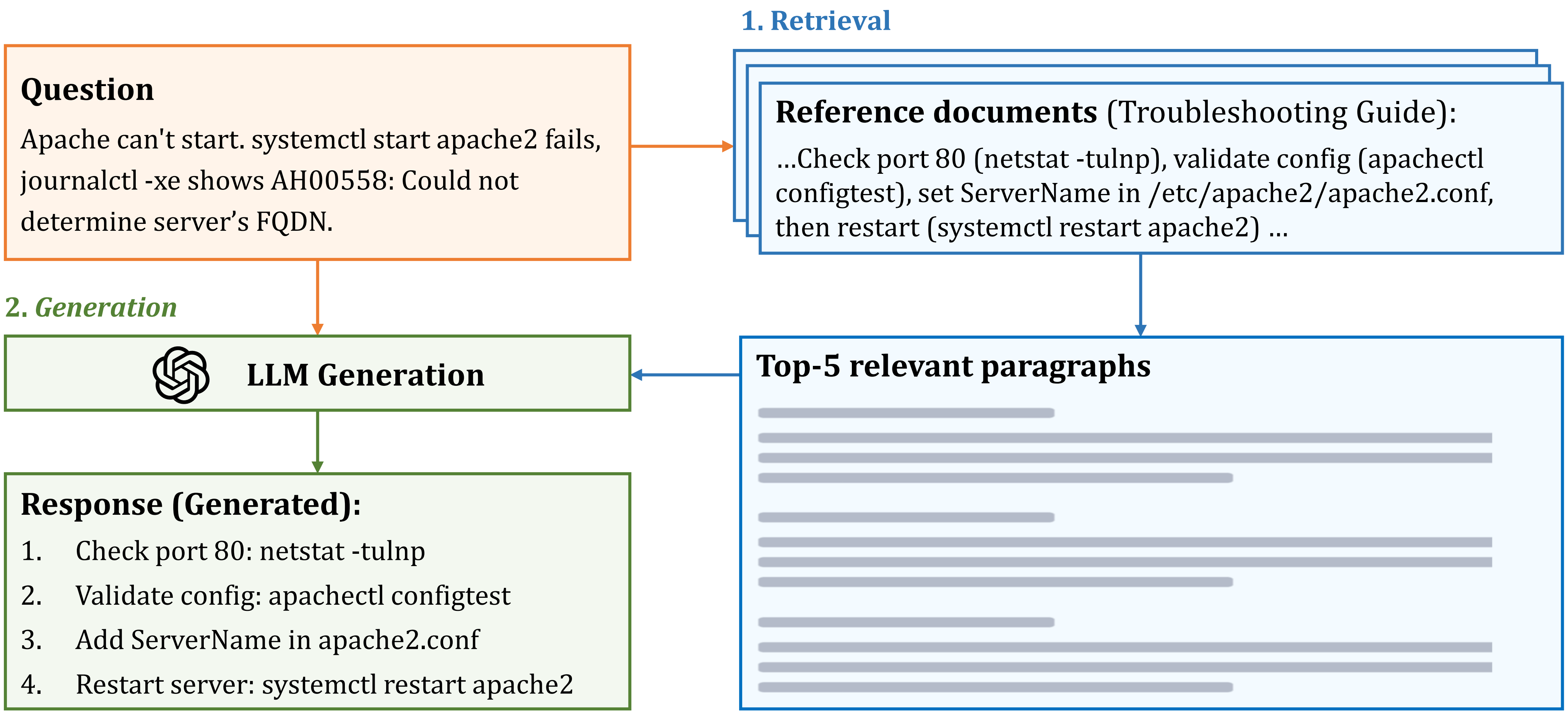
Example from Microsoft Forum

Question:
Are there ways to capture the IP addresses that are using my storage accounts in Azure portal?

Response:
Once you have enabled logging for your storage account, the information about operations performed against your storage account is saved in **`$logs` blob container**. It contains a CSV files. The information you're looking for is available in **<requester-ip-address> field**.

Background

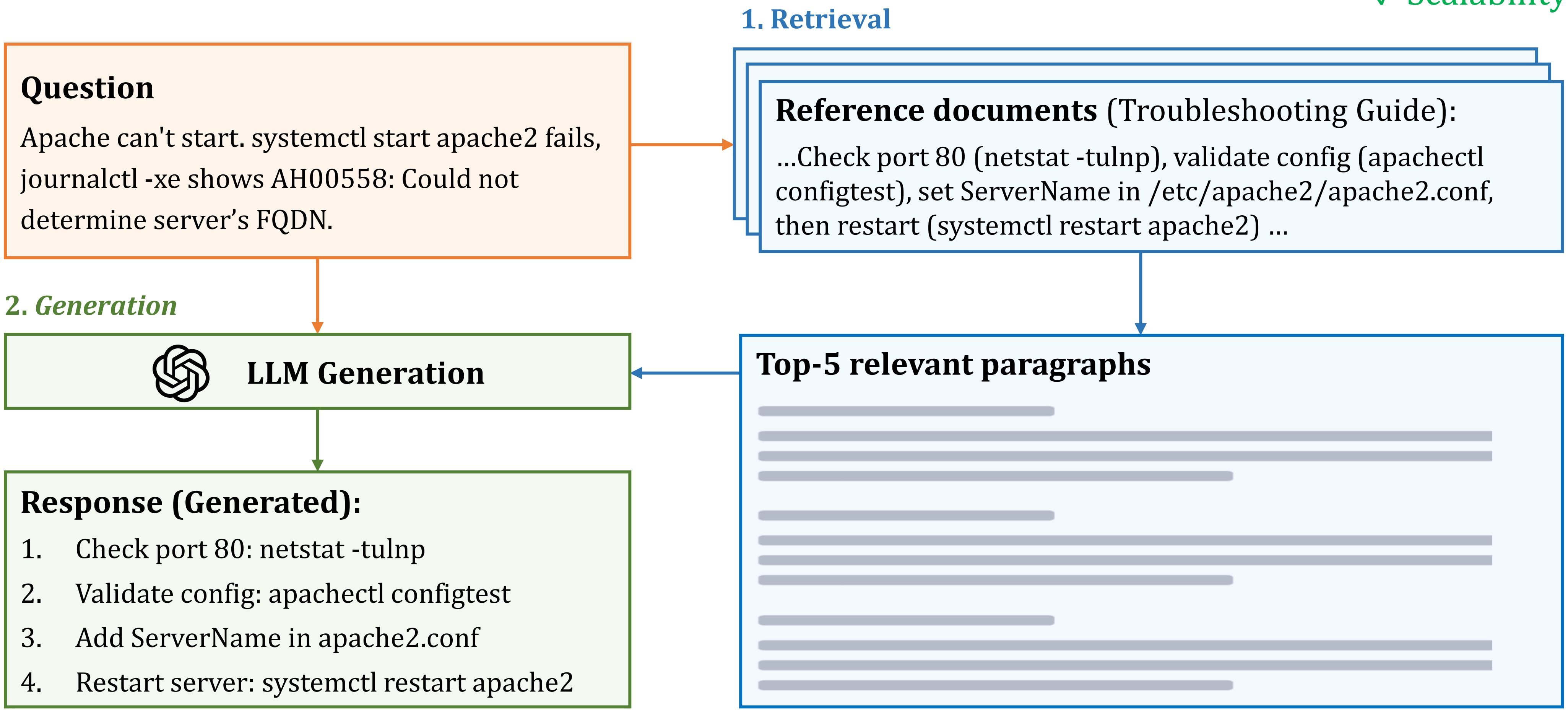
From Manual Responses to LLM-RAG Powered QA



Background

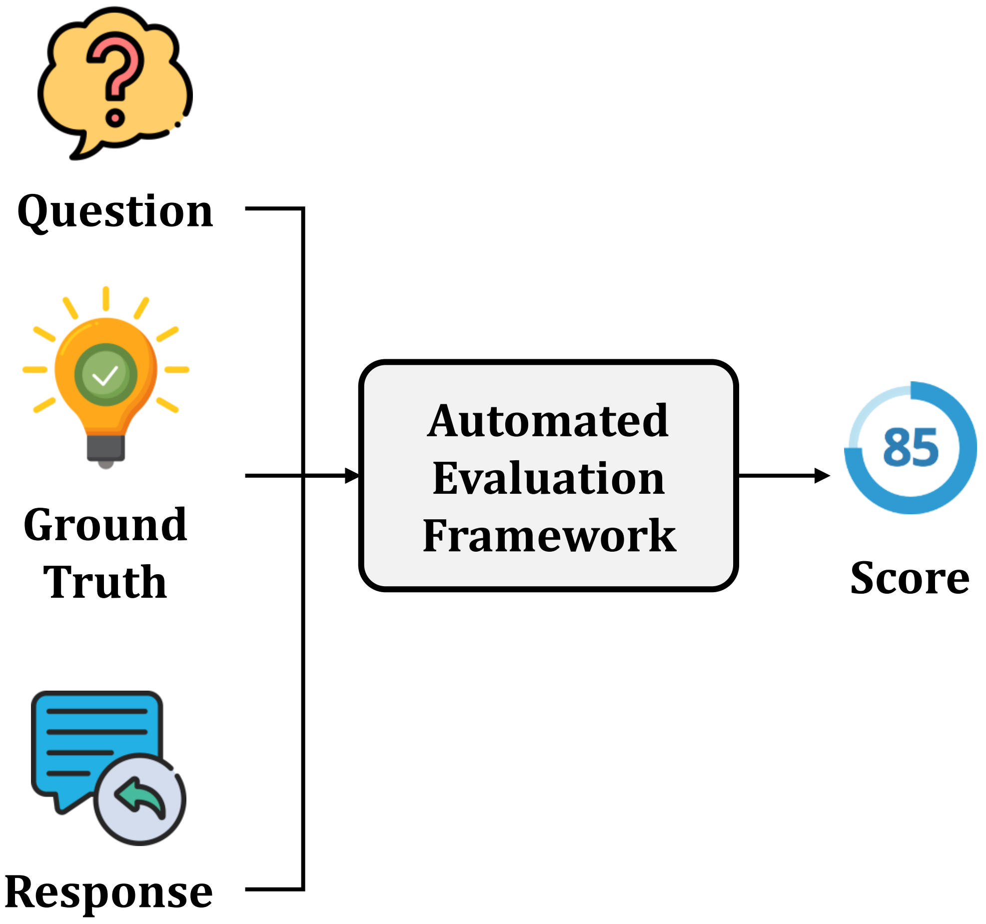
From Manual Responses to LLM-RAG Powered QA

? Accuracy
✓ Latency
✓ Scalability



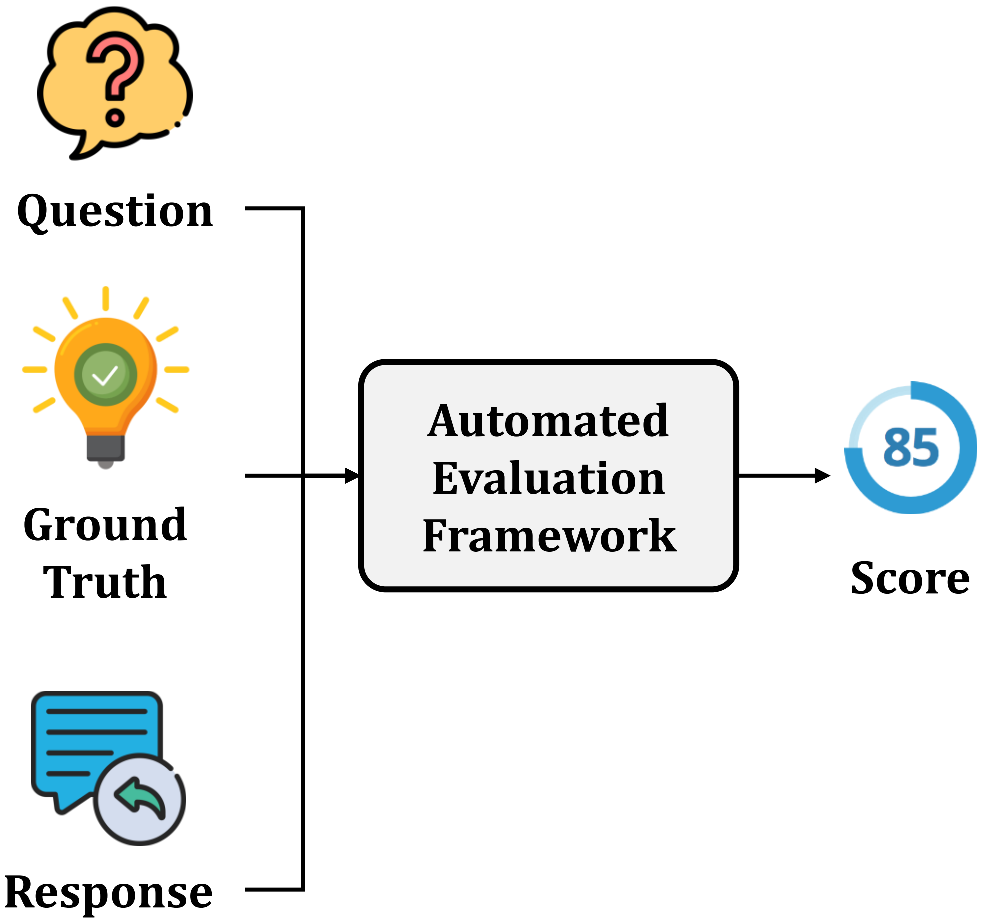
Background

Automated Evaluation of QA



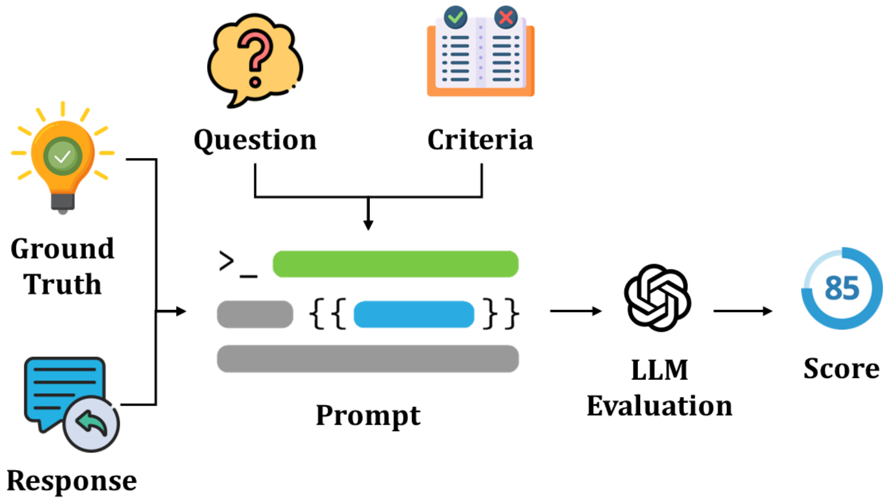
Background

Automated Evaluation of QA

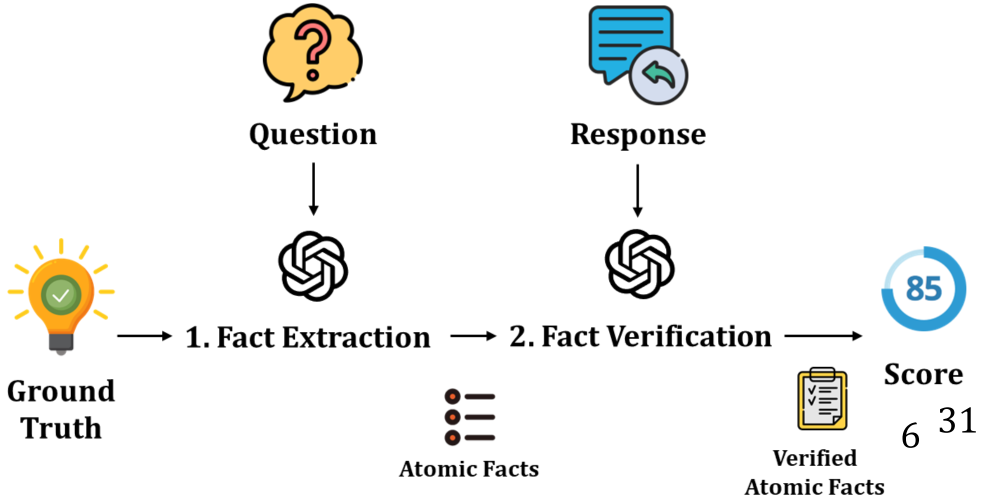


Existing evaluation methods:

1. Criteria-Guided Evaluation (e.g. G-Eval)



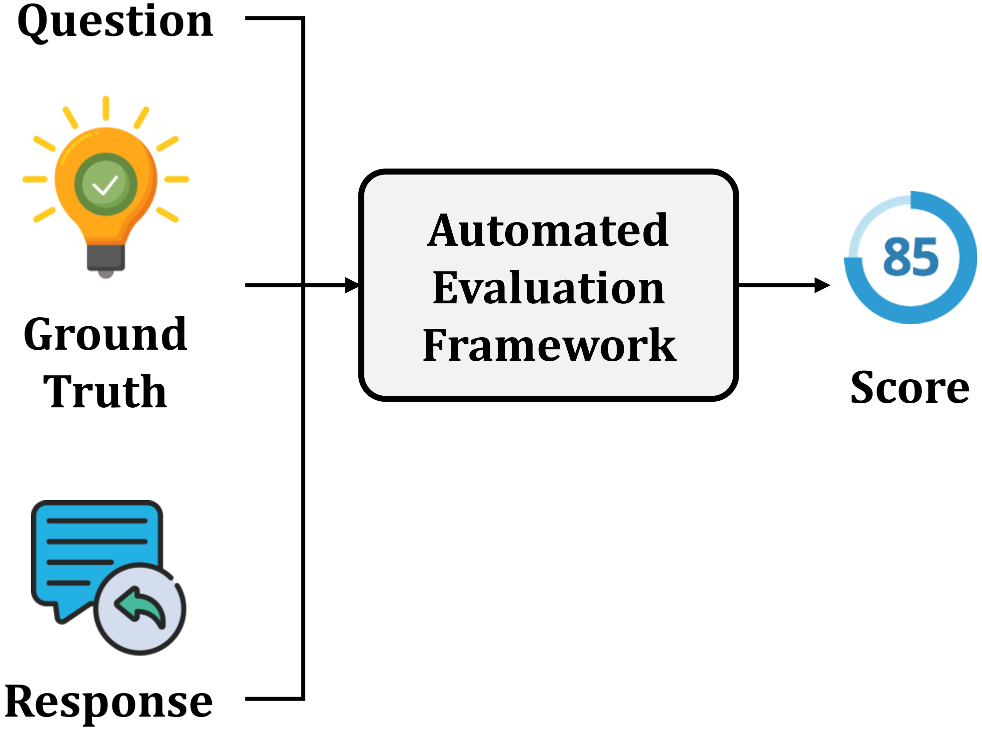
2. Factual Consistency Evaluation (e.g. RAGAS)



Background

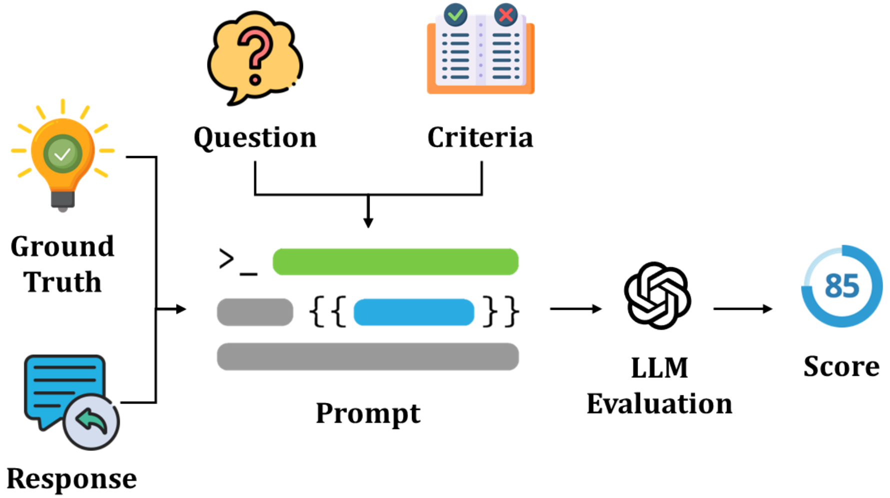
Automated Evaluation of QA

Observation:
Existing methods fail to capture **critical errors** in tech support QA.

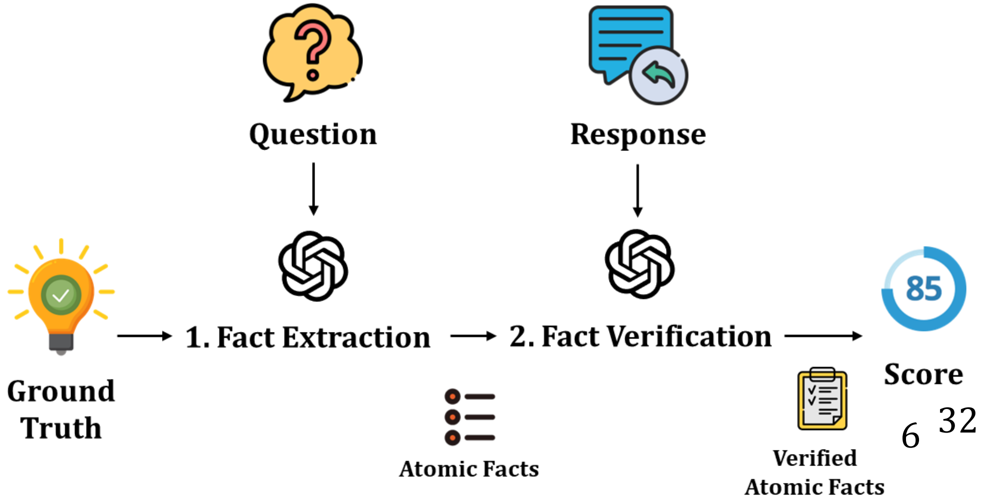


Existing evaluation methods:

1. Criteria-Guided Evaluation (e.g. G-Eval)



2. Factual Consistency Evaluation (e.g. RAGAS)



Background

Error Typology of Technical Support QA

Question *Q*

My Apache server fails to start. Running systemctl start apache2 shows an error. How can I fix this?

Ground Truth *GT*

1. Identify the process using port 80 with netstat -tulnp.
2. Stop the process.
3. Restart the server.

Error Type	Response *A*
Key Term Mismatch	1. Identify the process with netstat -anp. 2. Stop the process. 3. Restart the server.
Step Missing	1. Identify the process with netstat -tulnp. 2. Restart the server. (Missing step 2 in ground truth)
Step Reversal	1. Restart the server. (This should be the last step) 2. Identify the process with netstat -tulnp. 3. Stop the process.

Background

Challenge: Detecting the Critical Errors

Key Term Matching

- LLMs may hallucinate or omit key terms such as commands and file paths.
- These mistakes can mislead users and result in faulty or harmful operations.

Step Order Verification

- LLMs often fail to preserve the correct order in multi-step solutions.
- Incorrect step order may lead to configuration failures or system errors.

Step Completeness Verification

- RAG-based QA system tend to skip steps during retrieval.
- Missing steps result in incomplete guidance, leaving users unable to resolve the issue.

OUTLINE

Background

Framework

Evaluation

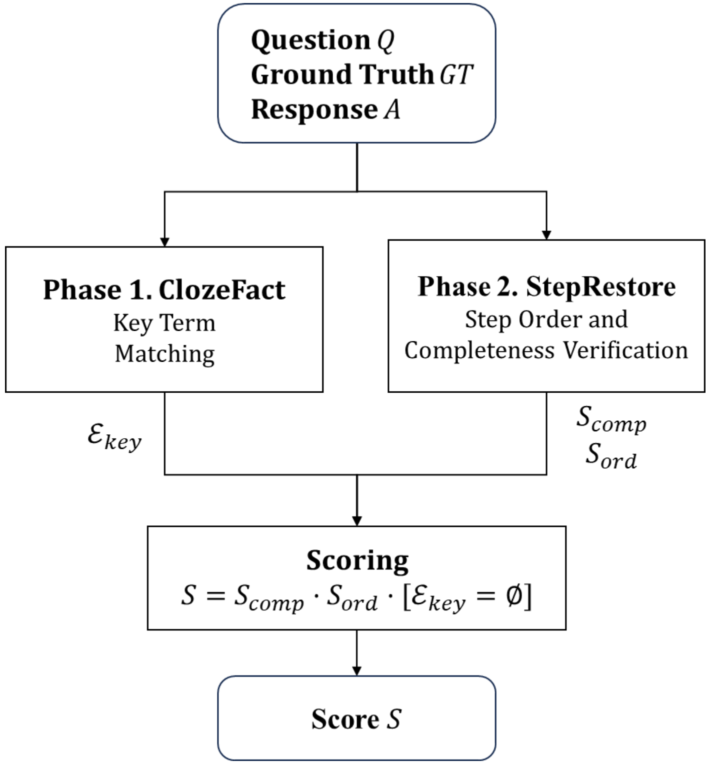
Conclusion

Framework

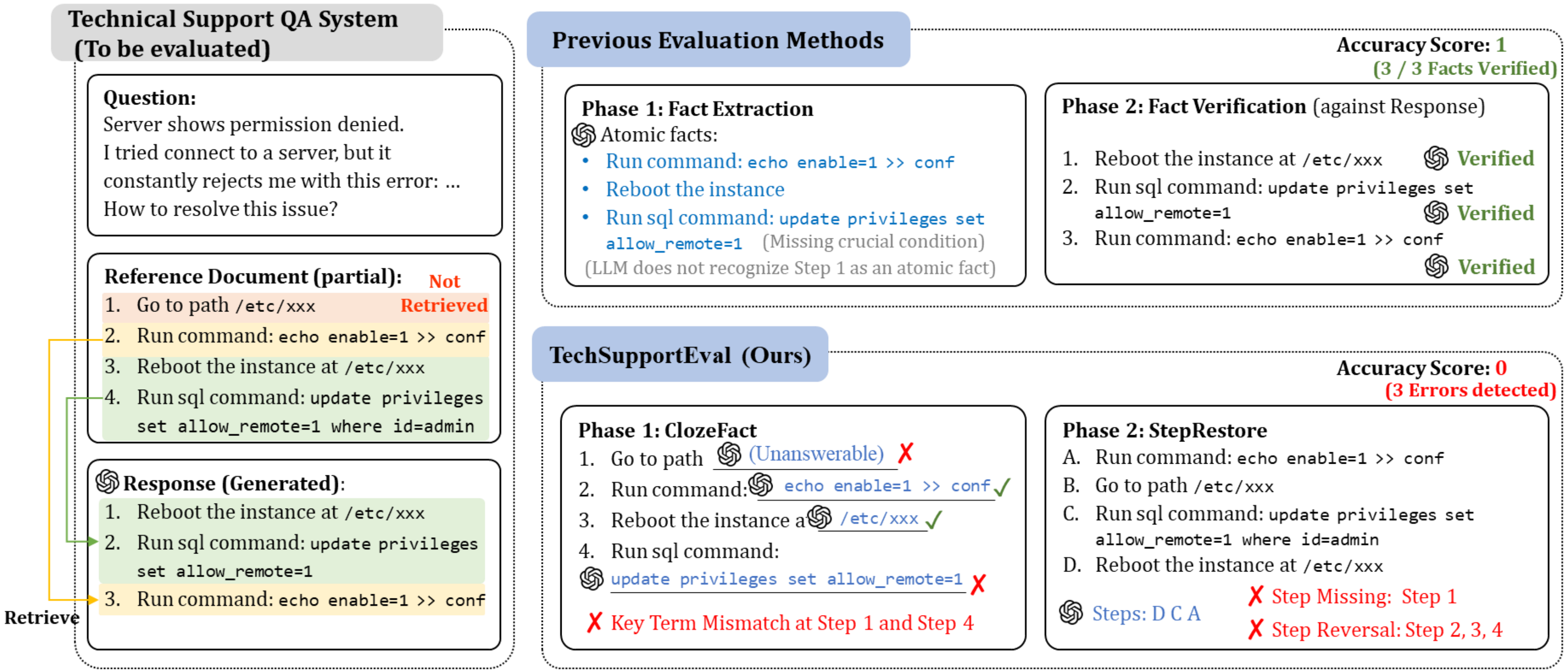
TechSupportEval: Overview

- An automated evaluation framework for technical support QA

TechSupportEval Workflow:



Comparison of TechSupportEval with previous evaluation methods:



Framework

Phase 1: ClozeFact

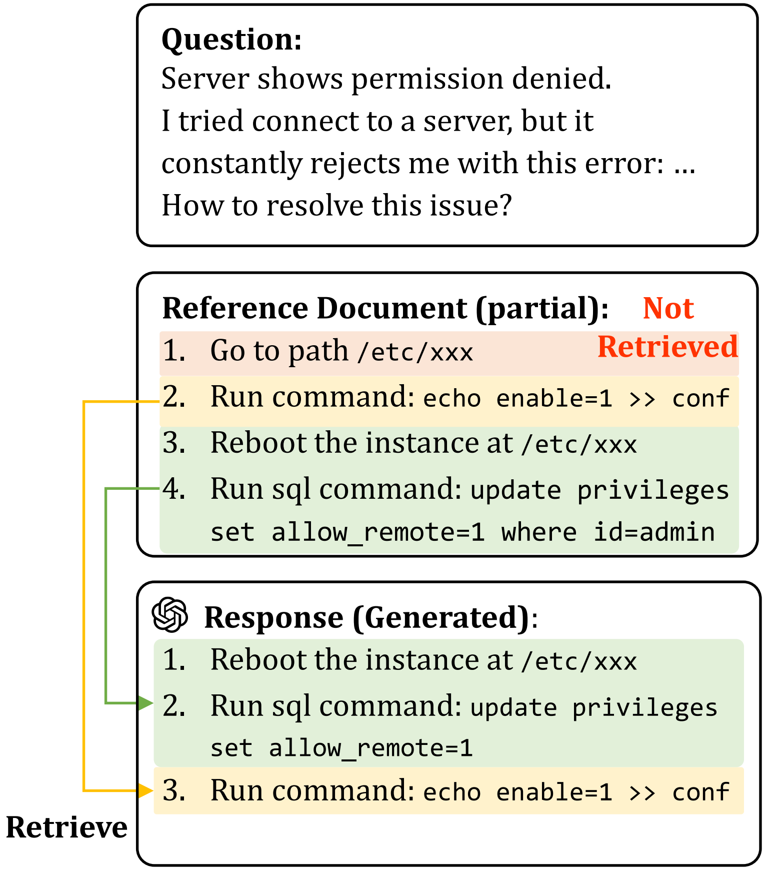
Question:
Server shows permission denied.
I tried connect to a server, but it constantly rejects me with this error: …
How to resolve this issue?

Reference Document (partial):

1. Go to path `/etc/xxx`
2. Run command: `echo enable=1 >> conf`
3. Reboot the instance at `/etc/xxx`
4. Run sql command: `update privileges set allow_remote=1 where id=admin`

Framework

Phase 1: ClozeFact

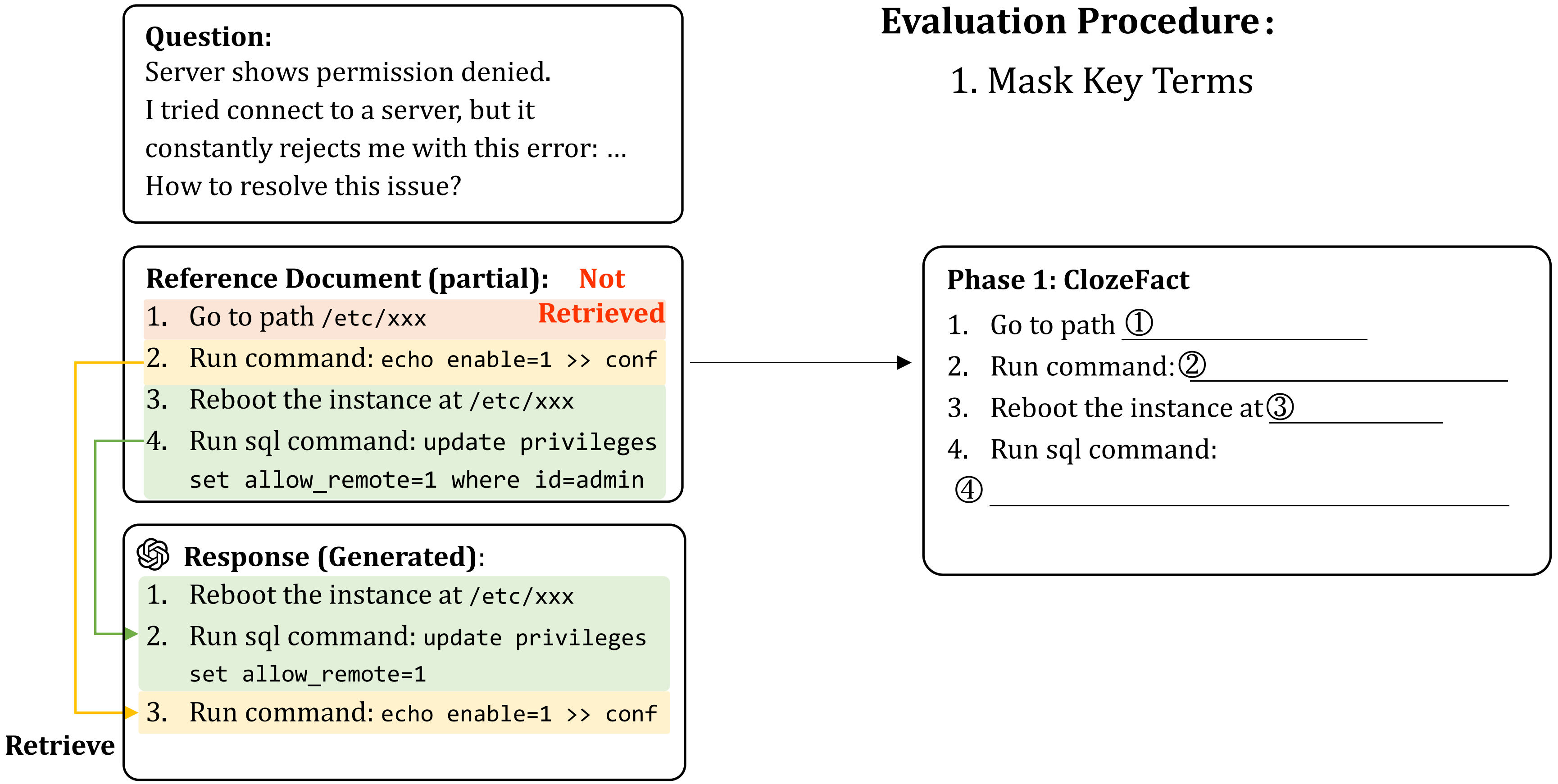


Framework

Phase 1: ClozeFact

Evaluation Procedure:

1. Mask Key Terms



Framework

Phase 1: ClozeFact

Question:
Server shows permission denied.
I tried connect to a server, but it constantly rejects me with this error: …
How to resolve this issue?

Reference Document (partial): Not Retrieved

1. Go to path `/etc/xxx`
2. Run command: `echo enable=1 >> conf`
3. Reboot the instance at `/etc/xxx`
4. Run sql command: `update privileges set allow_remote=1 where id=admin`

Response (Generated):

1. Reboot the instance at `/etc/xxx`
2. Run sql command: `update privileges set allow_remote=1`
3. Run command: `echo enable=1 >> conf`

Retrieve

Evaluation Procedure:

1. Mask Key Terms
2. Fill Blanks (using LLM_{eval})

Phase 1: ClozeFact

1. Go to path ① `(Unanswerable)`
2. Run command: ② `echo enable=1 >> conf`
3. Reboot the instance at ③ `/etc/xxx`
4. Run sql command: ④ `update privileges set allow_remote=1`

Framework

Phase 1: ClozeFact

Question:
Server shows permission denied.
I tried connect to a server, but it constantly rejects me with this error: ...
How to resolve this issue?

Reference Document (partial): Not Retrieved

1. Go to path `/etc/xxx`
2. Run command: `echo enable=1 >> conf`
3. Reboot the instance at `/etc/xxx`
4. Run sql command: `update privileges set allow_remote=1 where id=admin`

Response (Generated):

1. Reboot the instance at `/etc/xxx`
2. Run sql command: `update privileges set allow_remote=1`
3. Run command: `echo enable=1 >> conf`

Retrieve

Evaluation Procedure:

1. Mask Key Terms
2. Fill Blanks (using LLM_{eval})
3. Match Key Terms

Phase 1: ClozeFact

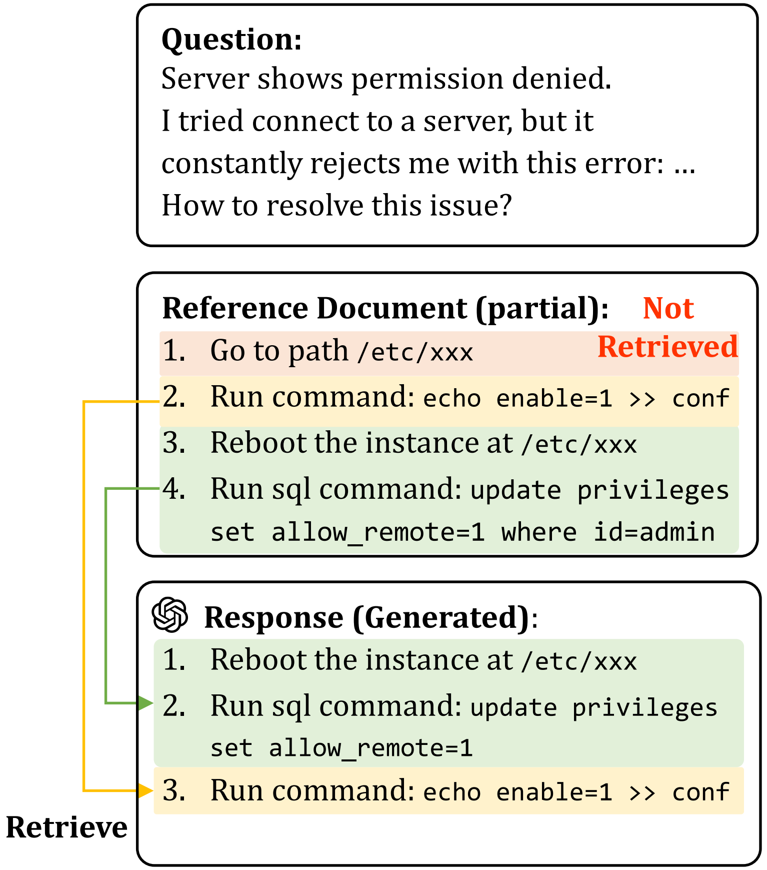
1. Go to path ① `(Unanswerable)` ✗
2. Run command: ② `echo enable=1 >> conf` ✓
3. Reboot the instance at ③ `/etc/xxx` ✓
4. Run sql command: ④ `update privileges set allow_remote=1` ✗

✗ Key Term Mismatch at ① and ④

$$\mathcal{E}_{key} = \{①, ④\}$$

Framework

Phase 2: StepRestore



Framework

Phase 2: StepRestore

Question:
Server shows permission denied.
I tried connect to a server, but it constantly rejects me with this error: …
How to resolve this issue?

Reference Document (partial): Not Retrieved

1. Go to path `/etc/xxx`
2. Run command: `echo enable=1 >> conf`
3. Reboot the instance at `/etc/xxx`
4. Run sql command: `update privileges set allow_remote=1 where id=admin`

Response (Generated):

1. Reboot the instance at `/etc/xxx`
2. Run sql command: `update privileges set allow_remote=1`
3. Run command: `echo enable=1 >> conf`

Retrieve

Evaluation Procedure:

1. Shuffle Steps

Phase 2: StepRestore

A. Run command: `echo enable=1 >> conf`
B. Go to path `/etc/xxx`
C. Run sql command: `update privileges set allow_remote=1 where id=admin`
D. Reboot the instance at `/etc/xxx`

Framework

Phase 2: StepRestore

Question:
Server shows permission denied.
I tried connect to a server, but it constantly rejects me with this error: ...
How to resolve this issue?

Reference Document (partial): **Not Retrieved**

1. Go to path `/etc/xxx`
2. Run command: `echo enable=1 >> conf`
3. Reboot the instance at `/etc/xxx`
4. Run sql command: `update privileges set allow_remote=1 where id=admin`

Response (Generated):

1. Reboot the instance at `/etc/xxx`
2. Run sql command: `update privileges set allow_remote=1`
3. Run command: `echo enable=1 >> conf`

Retrieve

Evaluation Procedure:

1. Shuffle Steps
2. Reorder Steps(using LLM_{eval})

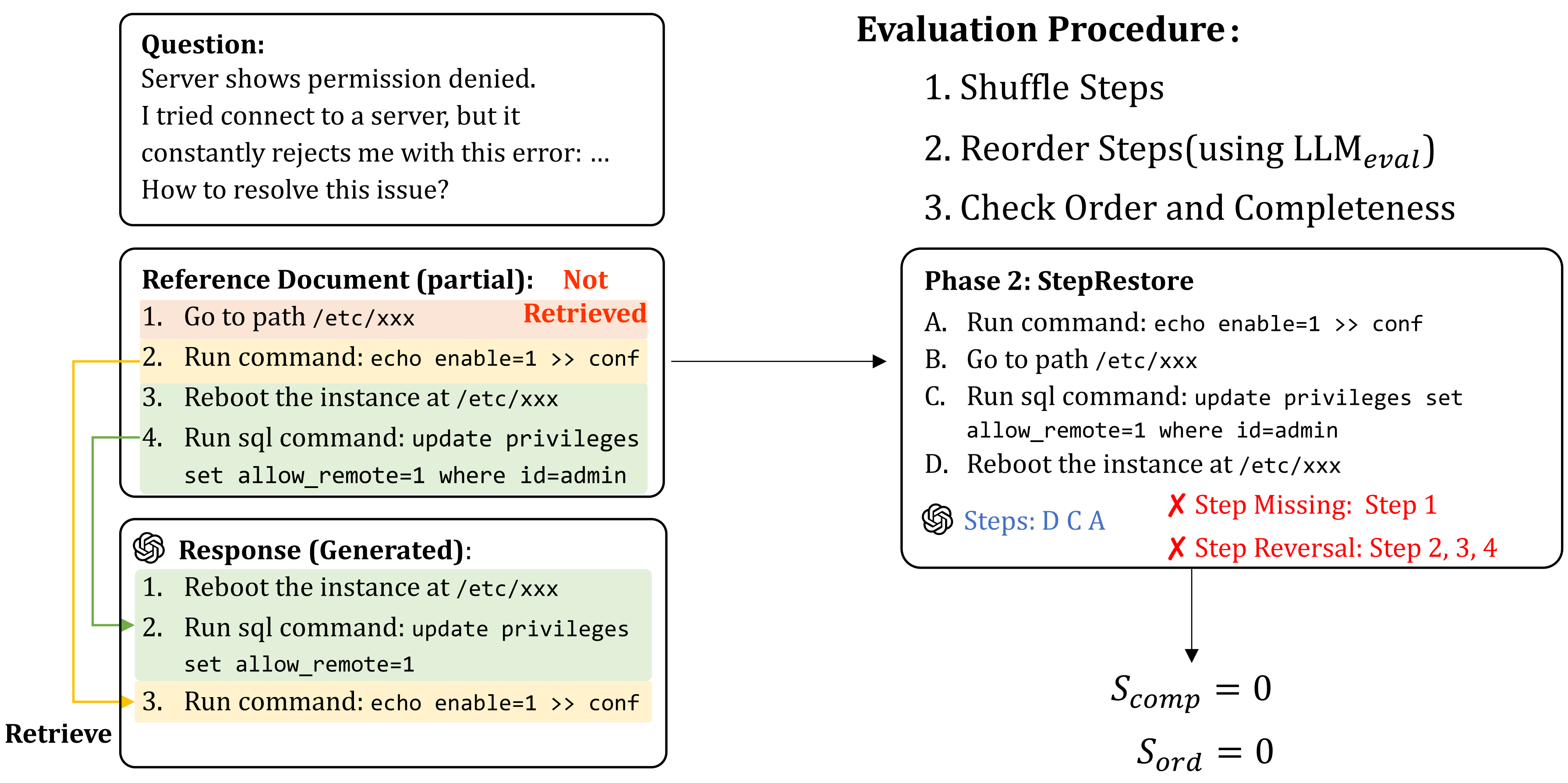
Phase 2: StepRestore

A. Run command: `echo enable=1 >> conf`
B. Go to path `/etc/xxx`
C. Run sql command: `update privileges set allow_remote=1 where id=admin`
D. Reboot the instance at `/etc/xxx`

Steps: D C A

Framework

Phase 2: StepRestore



Framework

Scoring Strategy

Phase 1.ClozeFact	Phase 2. StepRestore
Key Term Matching	Step Order and Completeness Verification
$\mathcal{E}_{key}$	S_{comp}, S_{ord}

- **Strict Scoring** (Default in TechSupportEval)
  - Reflects the strict requirement in technical support.
  - Any critical error—including incorrect key terms, missing steps, or wrong step order— would result in a failing score.

$$S = S_{comp} \cdot S_{ord} \cdot [\mathcal{E}_{key} = \varnothing]$$

- **Weighted Scoring** (Flexible Alternative)
  - Designed for scenarios with higher tolerance for minor issues.
  - A parameter α balances the impact of different error types.
  - Allows partial score when step order is correct, even if some steps are missing, aligning better with real-world user expectations.

$$S = \alpha \cdot S_{CF} + (1-\alpha) \cdot S_{SR}$$

$$S_{CF} = 1 - \frac{|\mathcal{E}_{key}|}{|K|}$$

$$S_{SR} = \frac{1}{2}(S_{comp} + S_{ord})$$

Framework

Implementation

Evaluation Workflow Management

- Automatically evaluates QA samples using a modular pipeline.
- Standardized process for generating and scoring answers.

Parallel Execution Strategy

- Runs two evaluation phases in parallel per sample.
- Supports sample-level parallelism for faster evaluation.

Unified LLM Interface

- Unified interface for both API-based and local LLMs.
- Built-in adaptive rate control for stable evaluation.

EvaluateOneSample($Q, GT, \text{LLM}_{QA}, \text{LLM}_{eval}$)

$A \leftarrow$ GenerateAnswer(Q, LLM_{QA})
in parallel do
    $\mathcal{E}_{key} \leftarrow$ ClozeFact(A, GT, LLM_{eval})
    $(S_{comp}, S_{ord}) \leftarrow$ StepRestore(A, GT, LLM_{eval})
wait until both modules complete
$S \leftarrow$ Scoring($\mathcal{E}_{key}, S_{comp}, S_{ord}$)
return S

function ClozeFact(A, GT, LLM_{eval})
    $K \leftarrow$ ExtractKeyTerms(GT)
    $GT' \leftarrow$ MaskKeyTerms(GT, K)
    $A' \leftarrow$ FillBlanks(GT', A, LLM_{eval})
    $\mathcal{E}_{key} \leftarrow$ MatchKeyTerms(A', K)
    return $\mathcal{E}_{key}$
end function

function StepRestore(A, GT, LLM_{eval})
    $GT'' \leftarrow$ ShuffleSteps(GT)
    $A_{rec} \leftarrow$ ReorderSteps(GT'', A, LLM_{eval})
    $S_{comp} \leftarrow$ CheckCompleteness(A_{rec}, GT)
    $S_{ord} \leftarrow$ CheckOrder(A_{rec}, GT)
    return (S_{comp}, S_{ord})
end function

OUTLINE

Background

Framework

Evaluation

Conclusion

Evaluation

Dataset

Our Dataset

TechQA[1] Dataset

Question Q

Ground Truth GT

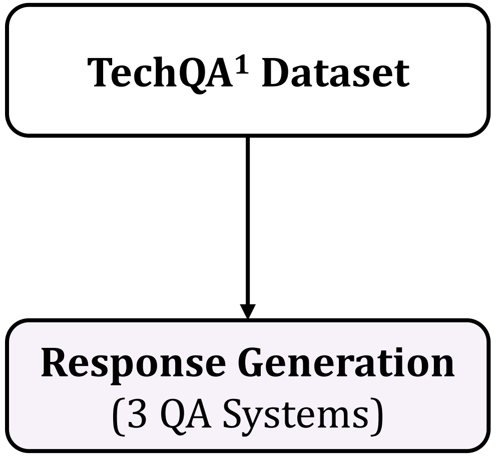
Metric	**Value**
Number of Questions	282
Avg. Length of Questions	366.48
Avg. Length of Ground Truths	220.87
Avg. Length of Reference Documents	4844.93
Avg. Steps in Ground Truths	2.04
Max. Steps in Ground Truths	14

Stats of the filtered TechQA Dataset

[1] Castelli et al. "The TechQA Dataset." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Evaluation

Dataset



Our Dataset

Question Q

Ground Truth GT

+

Response A

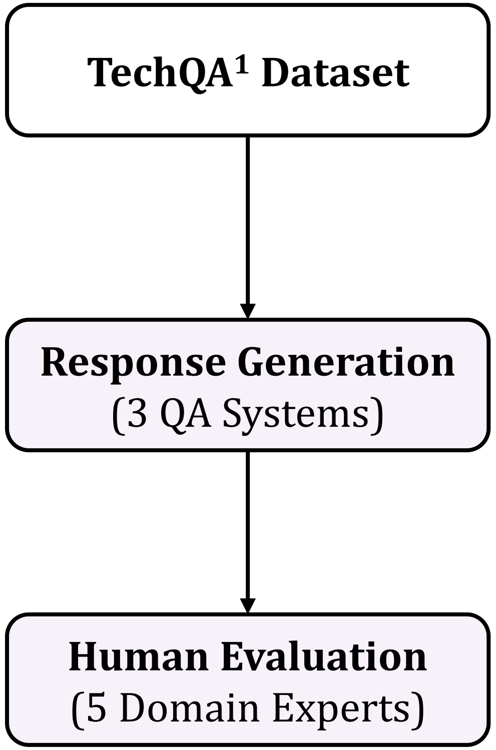
Metric	Value
Number of Questions	282
Avg. Length of Questions	366.48
Avg. Length of Ground Truths	220.87
Avg. Length of Reference Documents	4844.93
Avg. Steps in Ground Truths	2.04
Max. Steps in Ground Truths	14

Stats of the filtered TechQA Dataset

[1] Castelli et al. “The TechQA Dataset.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Evaluation

Dataset



Our Dataset

Question Q

Ground Truth GT

+

Response A

+

Human-Evaluated Score S^*

Metric	**Value**
Number of Questions	282
Avg. Length of Questions	366.48
Avg. Length of Ground Truths	220.87
Avg. Length of Reference Documents	4844.93
Avg. Steps in Ground Truths	2.04
Max. Steps in Ground Truths	14

Stats of the filtered TechQA Dataset

QA System (LLM_{QA})	**Accuracy**
GPT 4o Mini	0.8440
LLaMA 3 (70B)	0.7092
LLaMA 3 (8B)	0.5284

Human-Evaluated Accuracy $\overline{S^*}$ (of 3 QA Systems)

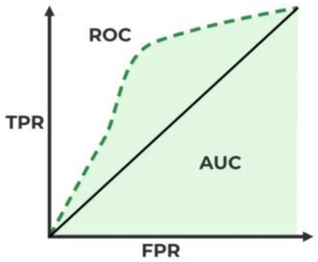
[1] Castelli et al. "The TechQA Dataset." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

Evaluation

Effectiveness & Ablation Study

Type	Method	LLM of Evaluated QA Systems					
		GPT 4o Mini		LLaMA 3 (70B)		LLaMA 3 (8B)	
		AUC	Pearson r	AUC	Pearson r	AUC	Pearson r
Lexical-based	ROUGE-1	0.5321	0.0311	0.5420	0.0648	0.5288	0.0484
	ROUGE-L	0.5631	0.0872	0.5932	0.1615	0.5752	0.1554
	BLEU	0.6061	0.1138	0.6252	0.1940	0.6158	0.1959
Semantic-based	BERTScore	0.6584	0.2243	0.6793	0.2892	0.6894	0.3095
LLM-based	LangChain Eval.	0.6608	0.4034	0.6310	0.3525	0.7015	0.4431
	LlamaIndex Eval.	0.6651	0.3061	0.6849	0.4117	0.7899	0.5131
	RAGAS	0.6728	0.1934	0.6894	0.2730	0.6544	0.2531
	RAGQuestEval	0.7416	0.3546	0.7205	0.3768	0.6899	0.3380
	G-Eval	0.8233	0.5192	0.8169	0.5419	0.8532	0.6109
	RefChecker	0.8348	0.4627	0.8313	0.5493	0.8309	0.5862
LLM-based	TECHSUPPORTEVAL	0.9109	**0.6616**	**0.8876**	**0.7430**	**0.8970**	**0.7938**
	w/o **ClozeFact**	0.8486	0.4641	0.8463	0.5752	0.8323	0.5914
	w/o **StepRestore**	**0.9129**	0.5669	0.8517	0.5884	0.8693	0.6635

AUC (Area Under the ROC Curve)



$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Pearson *r*

$$r = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(s_i^* - \bar{s^*})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2} \cdot \sqrt{\sum_{i=1}^{N}(s_i^* - \bar{s^*})^2}}$$

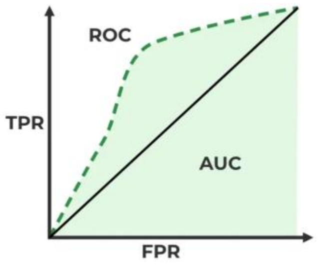
Evaluation

Effectiveness & Ablation Study

Type	Method	GPT 4o Mini AUC	GPT 4o Mini Pearson r	LLaMA 3 (70B) AUC	LLaMA 3 (70B) Pearson r	LLaMA 3 (8B) AUC	LLaMA 3 (8B) Pearson r
Lexical-based	ROUGE-1	0.5321	0.0311	0.5420	0.0648	0.5288	0.0484
	ROUGE-L	0.5631	0.0872	0.5932	0.1615	0.5752	0.1554
	BLEU	0.6061	0.1138	0.6252	0.1940	0.6158	0.1959
Semantic-based	BERTScore	0.6584	0.2243	0.6793	0.2892	0.6894	0.3095
LLM-based	LangChain Eval.	0.6608	0.4034	0.6310	0.3525	0.7015	0.4431
	LlamaIndex Eval.	0.6651	0.3061	0.6849	0.4117	0.7899	0.5131
	RAGAS	0.6728	0.1934	0.6894	0.2730	0.6544	0.2531
	RAGQuestEval	0.7416	0.3546	0.7205	0.3768	0.6899	0.3380
	G-Eval	0.8233	0.5192	0.8169	0.5419	0.8532	0.6109
	RefChecker	0.8348	0.4627	0.8313	0.5493	0.8309	0.5862
LLM-based	TechSupportEval	0.9109	**0.6616**	**0.8876**	**0.7430**	**0.8970**	**0.7938**
	w/o **ClozeFact**	0.8486	0.4641	0.8463	0.5752	0.8323	0.5914
	w/o **StepRestore**	**0.9129**	0.5669	0.8517	0.5884	0.8693	0.6635

TechSupportEval significantly **outperforms** all baseline methods in **evaluation accuracy.**

AUC (Area Under the ROC Curve)



$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Pearson *r*

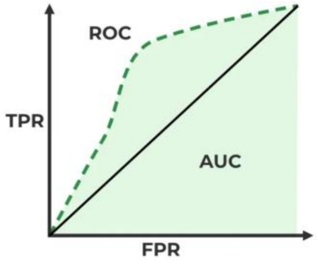
$$r = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(s_i^* - \bar{s^*})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2} \cdot \sqrt{\sum_{i=1}^{N}(s_i^* - \bar{s^*})^2}}$$

Evaluation

Effectiveness & Ablation Study

Type	Method	LLM of Evaluated QA Systems					
		GPT 4o Mini		LLaMA 3 (70B)		LLaMA 3 (8B)	
		AUC	Pearson r	AUC	Pearson r	AUC	Pearson r
Lexical-based	ROUGE-1	0.5321	0.0311	0.5420	0.0648	0.5288	0.0484
	ROUGE-L	0.5631	0.0872	0.5932	0.1615	0.5752	0.1554
	BLEU	0.6061	0.1138	0.6252	0.1940	0.6158	0.1959
Semantic-based	BERTScore	0.6584	0.2243	0.6793	0.2892	0.6894	0.3095
LLM-based	LangChain Eval.	0.6608	0.4034	0.6310	0.3525	0.7015	0.4431
	LlamaIndex Eval.	0.6651	0.3061	0.6849	0.4117	0.7899	0.5131
	RAGAS	0.6728	0.1934	0.6894	0.2730	0.6544	0.2531
	RAGQuestEval	0.7416	0.3546	0.7205	0.3768	0.6899	0.3380
	G-Eval	0.8233	0.5192	0.8169	0.5419	0.8532	0.6109
	RefChecker	0.8348	0.4627	0.8313	0.5493	0.8309	0.5862
LLM-based	TECHSUPPORTEVAL	0.9109	**0.6616**	**0.8876**	**0.7430**	**0.8970**	**0.7938**
	w/o **ClozeFact**	0.8486	0.4641	0.8463	0.5752	0.8323	0.5914
	w/o **StepRestore**	**0.9129**	0.5669	0.8517	0.5884	0.8693	0.6635

AUC (Area Under the ROC Curve)



$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

Pearson *r*

$$r = \frac{\sum_{i=1}^{N}(s_i - \bar{s})(s_i^* - \bar{s^*})}{\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2} \cdot \sqrt{\sum_{i=1}^{N}(s_i^* - \bar{s^*})^2}}$$

TechSupportEval significantly **outperforms** all baseline methods in **evaluation accuracy.**

Both ClozeFact and StepRestore **contribute to the overall performance.**

Evaluation

Impact of LLM$_{eval}$

LLM$_{eval}$	Method	LLM of Evaluated QA Systems					
		GPT 4o Mini		LLaMA 3 (70B)		LLaMA 3 (8B)	
		AUC	Pearson r	AUC	Pearson r	AUC	Pearson r
GPT 4o Mini	RAGAS	0.6728	0.1934	0.6544	0.2531	0.6894	0.2730
	RAGQuestEval	0.7416	0.3546	0.6899	0.3380	0.7205	0.3768
	RefChecker	0.8348	0.4627	0.8309	0.5862	0.8313	0.5493
	TECHSUPPORTEVAL	**0.9109**	**0.6616**	**0.8970**	**0.7938**	**0.8876**	**0.7430**
Claude 3.5 Haiku	RAGAS	0.7548	0.3489	0.7495	0.4285	0.7301	0.3767
	RAGQuestEval	0.7368	0.3337	0.7483	0.4319	0.7287	0.4197
	RefChecker	0.8132	0.4826	0.7704	0.4956	0.7681	**0.5391**
	TECHSUPPORTEVAL	**0.8651**	**0.5701**	**0.8029**	**0.6083**	**0.7737**	0.5332
LLaMA 3.3 70B	RAGAS	0.7705	0.3240	0.7579	0.4387	0.7066	0.3303
	RAGQuestEval	0.7807	0.3881	0.7056	0.3680	0.7394	0.4198
	RefChecker	0.8094	0.4533	0.7426	0.4256	0.7471	0.4492
	TECHSUPPORTEVAL	**0.8395**	**0.5021**	**0.8237**	**0.6464**	**0.7859**	**0.5538**
Qwen 2.5 72B	RAGAS	0.5199	0.0146	0.6919	0.3283	0.6017	0.1628
	RAGQuestEval	0.7342	0.3193	0.7481	0.4300	0.6746	0.2885
	RefChecker	0.7951	0.4041	0.7800	0.5033	0.7894	0.5339
	TECHSUPPORTEVAL	**0.8234**	**0.4954**	**0.8004**	**0.6013**	**0.8080**	**0.6002**

Evaluation

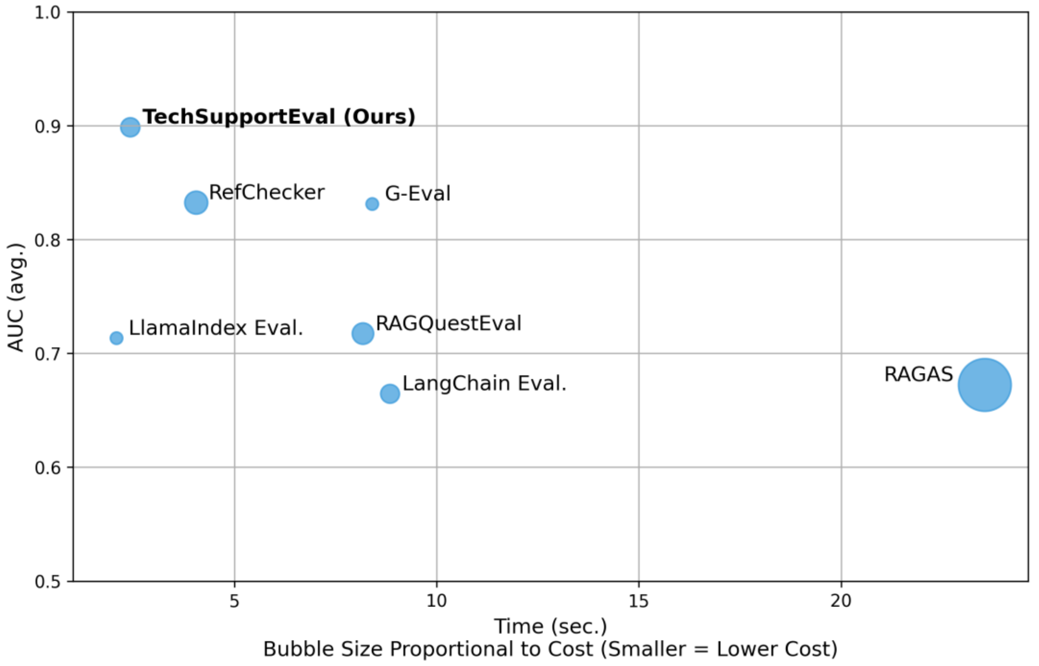
Impact of LLM$_{eval}$

LLM$_{eval}$	Method	LLM of Evaluated QA Systems					
		GPT 4o Mini		LLaMA 3 (70B)		LLaMA 3 (8B)	
		AUC	Pearson r	AUC	Pearson r	AUC	Pearson r
GPT 4o Mini	RAGAS	0.6728	0.1934	0.6544	0.2531	0.6894	0.2730
	RAGQuestEval	0.7416	0.3546	0.6899	0.3380	0.7205	0.3768
	RefChecker	0.8348	0.4627	0.8309	0.5862	0.8313	0.5493
	TECHSUPPORTEVAL	**0.9109**	**0.6616**	**0.8970**	**0.7938**	**0.8876**	**0.7430**
Claude 3.5 Haiku	RAGAS	0.7548	0.3489	0.7495	0.4285	0.7301	0.3767
	RAGQuestEval	0.7368	0.3337	0.7483	0.4319	0.7287	0.4197
	RefChecker	0.8132	0.4826	0.7704	0.4956	0.7681	**0.5391**
	TECHSUPPORTEVAL	**0.8651**	**0.5701**	**0.8029**	**0.6083**	**0.7737**	0.5332
LLaMA 3.3 70B	RAGAS	0.7705	0.3240	0.7579	0.4387	0.7066	0.3303
	RAGQuestEval	0.7807	0.3881	0.7056	0.3680	0.7394	0.4198
	RefChecker	0.8094	0.4533	0.7426	0.4256	0.7471	0.4492
	TECHSUPPORTEVAL	**0.8395**	**0.5021**	**0.8237**	**0.6464**	**0.7859**	**0.5538**
Qwen 2.5 72B	RAGAS	0.5199	0.0146	0.6919	0.3283	0.6017	0.1628
	RAGQuestEval	0.7342	0.3193	0.7481	0.4300	0.6746	0.2885
	RefChecker	0.7951	0.4041	0.7800	0.5033	0.7894	0.5339
	TECHSUPPORTEVAL	**0.8234**	**0.4954**	**0.8004**	**0.6013**	**0.8080**	**0.6002**

Robust across different backbone LLMs

Evaluation

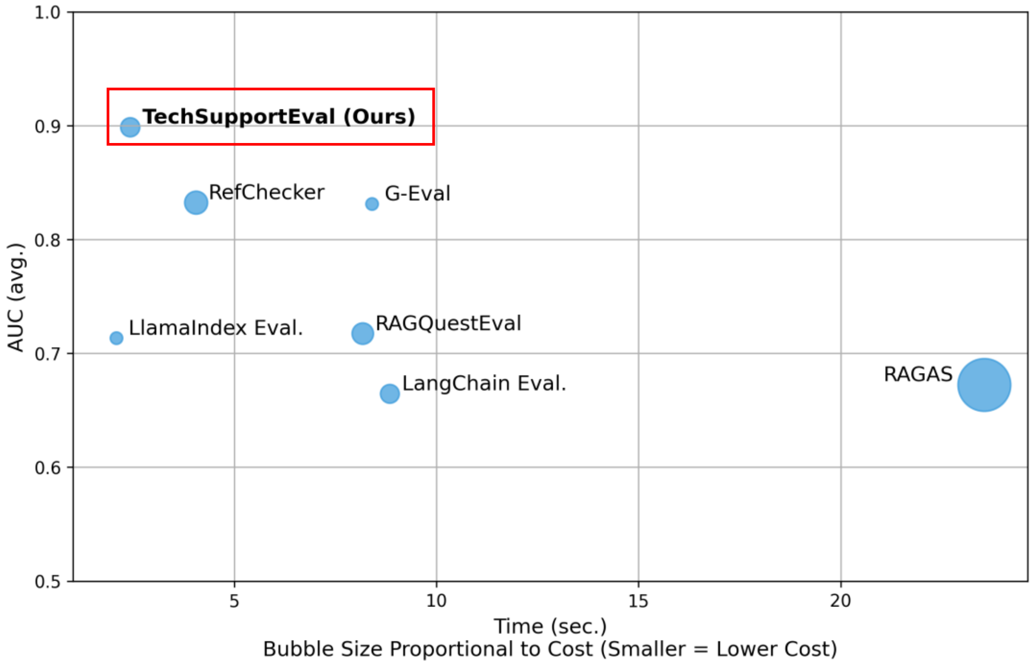
Efficiency and Cost



Method	**AUC (avg.)**	**Time (sec.)**	**Cost** ($10^{-3}\$$)
LangChain Eval.	0.6644	8.85	0.30
RAGAS	0.6722	23.55	2.37
LlamaIndex Eval.	0.7133	**2.09**	**0.13**
RAGQuestEval	0.7173	8.18	0.39
G-Eval	0.8311	8.41	**0.13**
RefChecker	0.8323	4.06	0.45
TECHSUPPORTEVAL	**0.8985**	2.43	0.31

Evaluation

Efficiency and Cost



Method	AUC (avg.)	Time (sec.)	Cost (10^{-3}\$)
LangChain Eval.	0.6644	8.85	0.30
RAGAS	0.6722	23.55	2.37
LlamaIndex Eval.	0.7133	**2.09**	**0.13**
RAGQuestEval	0.7173	8.18	0.39
G-Eval	0.8311	8.41	**0.13**
RefChecker	0.8323	4.06	0.45
TECHSUPPORTEVAL	**0.8985**	2.43	0.31

Efficient and **scalable** for large-scale evaluation

OUTLINE

Background

Framework

Evaluation

Conclusion

Conclusion

- We investigate **the evaluation of technical support QA** and pinpoint **three key challenges** it presents.
- We propose an LLM-based automated evaluation framework **TechSupportEval** for technical support QA with two novel techniques, **ClozeFact** and **StepRestore**, to address the challenges effectively.
- We **introduce a benchmark dataset** based on the publicly available TechQA dataset. Our approach achieves an **AUC of 0.91**, outperforming the previous state-of-the-art method by **7.6%**. The code and dataset are available at **https://github.com/NetManAIOps/TechSupportEval**.

Conclusion

- We investigate **the evaluation of technical support QA** and pinpoint **three key challenges** it presents.
- We propose an LLM-based automated evaluation framework **TechSupportEval** for technical support QA with two novel techniques, **ClozeFact** and **StepRestore**, to address the challenges effectively.
- We **introduce a benchmark dataset** based on the publicly available TechQA dataset. Our approach achieves an **AUC of 0.91**, outperforming the previous state-of-the-art method by **7.6%**. The code and dataset are available at **https://github.com/NetManAIOps/TechSupportEval**.

Conclusion

- We investigate **the evaluation of technical support QA** and pinpoint **three key challenges** it presents.
- We propose an LLM-based automated evaluation framework **TechSupportEval** for technical support QA with two novel techniques, **ClozeFact** and **StepRestore**, to address the challenges effectively.
- We **introduce a benchmark dataset** based on the publicly available TechQA dataset. Our approach achieves an **AUC of 0.91**, outperforming the previous state-of-the-art method by **7.6%**. The code and dataset are available at **https://github.com/NetManAIOps/TechSupportEval**.

Thank you!

Presenter: Bohan Chen

IJCNN 2025