Probing the Depths of Language Model's Contact Center Knowledge for Quality Assurance

Digvijay Ingle[1], Aashraya Sachdeva[1], Surya Prakash Sahu[2], Mayank Sati[2], Cijo George, Jithendra Vepa

OBSERVE·AI

EMNLP 2024

Empowering contact-centers to drive consistent, scalable and smarter agent evaluations

Motivation

- **Evaluating contact-center agents**
  - Requires deep domain knowledge & industry best-practices
  - Existing LMs struggle with nuances of QA (e.g. active listening, probing questions)
  - Unexplored use of LMs for agent-evaluation

➢ **Potential to improve – objectivity, consistency and scalability of QA**

Contact Center Knowledge of LMs

Given a question $\mathcal{Q}$, and conversation $\mathcal{C}$, we prompt an LM $\mathcal{L}$ to follow a chain-of-thought (CoT) reasoning:

1. Identify evidences relevant to $\mathcal{Q}$ in conversation $\mathcal{C}$
2. Generate synthesized reasoning based on the evidences
3. Conclude final answer $\mathcal{A}$

Distilling Domain-Knowledge to Small LMs

- **Approach 1:** Inference with Large LM Guided Plan
  - ➢ Use Large LM ($\mathcal{M}_{Sonnet}$) proficient in contact-center domain knowledge to generate evaluation plan $\mathcal{P}$
  - ➢ Prompt Small LM ($\mathcal{M}_{Phi}$) to follow CoT reasoning
- **Approach 2:** Fine-tuning with Large LM Generated Response
  - ➢ Generate a silver data with reasoning generated by a Large LM ($\mathcal{M}_{Sonnet}$)
  - ➢ Fine-tune small LM ($\mathcal{M}_{Phi}$) to follow CoT reasoning

> **Avoid interrupting the customer**: The agent avoided interrupting the customer while they were speaking, allowing them to fully explain their issue or concern.
> **Acknowledging customer's concerns**: The agent acknowledged or addressed the customer's concerns or questions.
> **Providing relevant responses**: The agent provided responses that were relevant and addressed the customer's actual issue or concern.

Figure 1: Evaluation Plan to assess an agent on: Did the agent demonstrate active listening?

> Did the agent demonstrate active listening?

> Did the agent address any objections raised by the customer?

> Did the agent ask probing questions to discover the customer's needs?

> Did the agent properly acknowledge customer inquiry?

Figure 1: QA Questions - Illustrative examples

Results

Group	Model	Macro F1 (%)
Large	GPT-4o	70.56
	Claude-3.5-sonnet	**75.48**
	Llama3-70B	74.68
Medium	GPT-4o-mini	72.97
	Llama3-8B	68.54
	Mistral-7B	62.96
Small	Phi-3-mini-128k-instruct	62.91
	Gemma-2B-it	54.17

Table 1: Strong correlation between model size and contact-center domain-knowledge

Setup	Fine-Tuned	Input	Output		Macro F1 (%)
		Plan	Evidence	Synthesis	
S0			✓	✓	62.91
S1		✓	✓	✓	67.99
S2	✓	✓	✓	✓	**81.86**
S2a	✓		✓	✓	78.80
S2b	✓	✓		✓	81.58

Table 2: Evaluation plan play a crucial role in distilling domain knowledge to smaller LMs

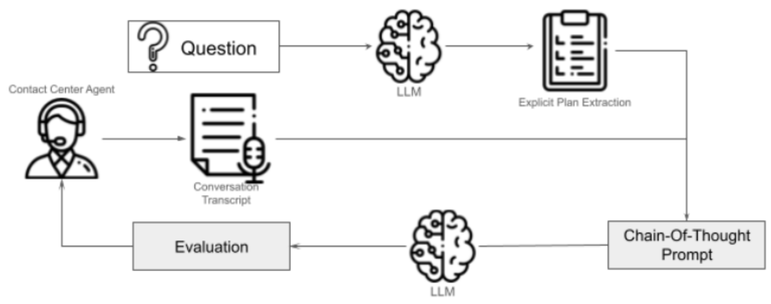

Figure 3: Human-in-the-loop QA process

Conclusion

- Model size highly correlates with contact-center domain-knowledge of LM's
- Evaluation plan generated by Large LM plays a crucial role in enhancing reasoning ability of small LMs

References

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart M. Shieber. Implicit chain of thought reasoning via knowledge distillation, 2023