Probing the Depths of Language Models' Contact-Center Knowledge for Quality Assurance

Digvijay Ingle[†] **Aashraya Sachdeva**[†] **Surya Prakash Sahu**[‡] **Mayank Sati**[‡]
Cijo George **Jithendra Vepa**
Observe.AI, India
{digvijay.ingle, aashraya.sachdeva, suryaprakash.sahu, mayank.sati}@observe.ai
{cijo.george, jithendra}@observe.ai

Abstract

Recent advancements in large Language Models (LMs) have significantly enhanced their capabilities across various domains, including natural language understanding and generation. In this paper, we investigate the application of LMs to the specialized task of contact-center Quality Assurance (QA), which involves evaluating conversations between human agents and customers. This task requires both sophisticated linguistic understanding and deep domain knowledge. We conduct a comprehensive assessment of eight LMs, revealing that larger models, such as Claude-3.5-Sonnet, exhibit superior performance in comprehending contact-center conversations. We introduce methodologies to transfer this domain-specific knowledge to smaller models by leveraging evaluation plans generated by more knowledgeable models, with optional human-in-the-loop refinement to enhance the capabilities of smaller models. Notably, our experimental results demonstrate an improvement of up to 18.95% in Macro F1 on an in-house QA dataset. Our findings emphasize the importance of evaluation plans in guiding reasoning and highlight the potential of AI-assisted tools to advance objective, consistent, and scalable agent evaluation processes in contact centers.

1 Introduction

The convergence of contact-center management and artificial intelligence represents a frontier rich with potential for revolutionizing customer service quality and operational efficiency. Contact-centers, serving as the primary interface between organizations and their customers, are increasingly seeking sophisticated methods to evaluate and enhance agent performance so as to improve their customer satisfaction (Roy et al., 2016). Concurrently, the field of natural language processing has witnessed unprecedented advancements with the emergence of large Language Models (LMs) such as GPT-4 (OpenAI, 2023), Gemini (Anil et al., 2023), Claude (Anthropic, 2023), and their successors. These models have demonstrated remarkable proficiency in understanding and generating human-like text across diverse domains, suggesting promising applications in a variety of natural language processing tasks, such as, machine translation (Zhu et al., 2024), sentiment analysis (Zhang et al., 2023), text summarisation (Van Veen et al., 2024; Yuan et al., 2024), reasoning (Wei et al., 2022b), etc.

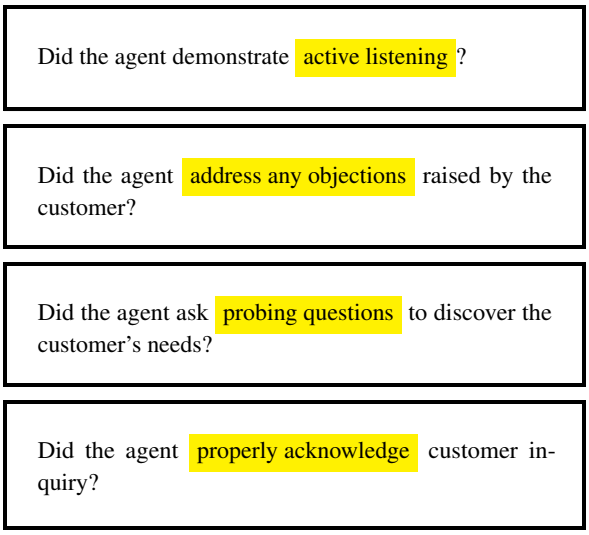

Figure 1: Real-world examples of QA questions for evaluating contact-center agents. Highlighted phrases indicate the domain-knowledge of contact-centers required to answer the respective QA questions.

However, evaluating contact-center agents using these LMs presents unique challenges that extend beyond basic linguistic comprehension. Effective assessment requires a deep understanding of industry best practices, domain knowledge, and the nuances of customer service communication. Let us consider the question - *"Did the agent demonstrate*

[†] Equal contribution as first authors.
[‡] Equal contribution as second authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 790–804

active listening?" illustrated in Figure 1. This evaluation involves more than just analyzing text; it requires a comprehensive grasp of active listening principles in customer interactions. An accurate assessment must determine if the agent attentively listened without requiring repetition, understood the customer's issue, and asked appropriate follow-up questions to guide the interaction towards resolution. While some aspects of this evaluation can be explicitly derived from the question, others demand deeper domain understanding (refer Appendix D). This includes recognizing the complexities of customer issues, appropriate troubleshooting steps, and the flow of effective customer service interactions. The multifaceted nature of this task highlights the need to integrate sophisticated NLP models with domain-specific expertise for comprehensive contact-center agent evaluations.

While research has explored LMs in various contact-center applications, their use in quality assurance remains understudied. Nathan et al. (2024) examine in-domain fine-tuning for tasks like summarization and question-answering, but does not address holistic agent evaluation. In this paper, we aim to fill this gap with three key contributions:

1. A comprehensive evaluation of eight LMs' ability to comprehend contact-center conversations for quality assurance purposes

2. Proposed methodologies for transferring domain knowledge to models lacking it, discussing practical implications

3. Future directions for developing AI-assisted evaluation tools in contact-centers, potentially enhancing objectivity, consistency, and scalability of assessments

2 Problem Formulation

Contact-centers typically have a dedicated quality assurance (QA) team responsible for maintaining high service standards and ensuring customer satisfaction. This team systematically evaluates agent performance across various interactions, focusing on adherence to company policies, compliance requirements, agent behaviour and best practices. As a part of this process, QA analysts meticulously review the agent-customer conversations, identify key events, and evaluate the agent's performance against predefined criteria. Maintaining consistency and accuracy of these evaluations poses a significant cognitive overload for QA analysts and is in turn a time-taking process, necessitating a nuanced approach to improve the efficiency and effectiveness of QA processes.

The evaluation criteria used by QA teams are often framed as questions, which need to be answered based on the conversation and the effectiveness of the agent's interaction with the customer. These questions cover various aspects of the interaction, such as whether the agent actively listened to the customer, accurately identified and addressed the issue, adhered to the company's communication protocols, etc. (refer to examples of QA questions in Figure 1). Framing the QA evaluation in this manner naturally makes it a question-answering task.

Importantly, providing just the answer to these questions is often not sufficient. Detailed reasoning must accompany each answer to validate the response and offer transparency. This reasoning explains why a particular answer was chosen, highlighting the relevant portions of the conversation that led to the conclusion (refer to Figure 4 in Appendix A.1 for a sample response to a QA question). This not only enhances the accuracy of the evaluation but also streamlines the process for QA professionals by offering clear justifications for each assessment, making their workflow more efficient and decisions more reliable.

Formally, we define the QA task as follows: Given a conversation $\mathcal{C}$ between an agent and a customer, and an evaluation question $\mathcal{Q}$ designed to assess a specific aspect of the agent's performance, the goal is to generate a detailed reasoning $\mathcal{R}$ and an appropriate answer $\mathcal{A}$, such that $\mathcal{R}$ logically leads to $\mathcal{A}$. This requires the extraction and synthesis of relevant information from the conversation $\mathcal{C}$, demonstrating a deep understanding of both contact-center domain-knowledge and nuances of the interaction. Providing $\mathcal{R}$ with $\mathcal{A}$ not only validates the response but also offers transparency and clarity, aiding QA analysts in their decision-making process.

3 Quantifying Contact-Center Knowledge of LMs

This section aims to evaluate out-of-the-box performance of a suite of language models (LMs) in the specific context of quality assurance (QA) within contact-centers. By benchmarking these LMs on their ability to answer the QA questions, we seek to understand the extent to which they can effectively

evaluate contact-center agents based on conversation transcripts, given their current knowledge of contact-center domains. The detailed methodology is outlined as follows:

3.1 Data Curation

To quantify the domain-knowledge of LMs, we curate a specialised quality assurance (QA) dataset. Specifically, we use a sample of 100 English dyadic conversations between agents and customers, transcribed using a third-party Automatic Speech Recognition (ASR) engine with a Word Error Rate (WER) (Ali and Renals, 2018) of 10%. We further sample a set of 50 QA questions from a proprietary contact-center dataset designed to holistically evaluate the performance of agents in handling customer interactions. Each of these questions can be answered as either `yes` or `no`. We then employ a group of seven annotators, who are experts in contact-center quality assurance to answer each of the 50 QA questions based on the 100 conversations. The annotators are provided with a comprehensive guideline to follow logical reasoning steps to identify relevant evidence from the conversation, synthesize them, and finally conclude the answer to the question. This approach not only ensures that the annotations are grounded in specific details from the conversations but also emulates the reasoning process a QA analyst would implicitly follow. To ensure the reliability of the dataset, we select question-conversation pairs where at least five annotators agree on the answer, resulting in a refined dataset of 3,061 question-conversation pairs with their annotated reasoning (evidence along with synthesis) and answer. This implies approximately 60% (3,061 out of 5,000) agreement between annotators. The answer agreed upon by the five annotators is selected as the ground-truth answer. For the ground truth reasoning, we first filter the reasonings corresponding to the selected ground truth answer and randomly sample one of those as the ground truth reasoning. This randomly selected reasoning is then post-processed to represent a coherent chain of thought that leads to the final answer, reflecting the logical steps followed by the annotators (see Appendix A.1 for annotated examples). We refer to this dataset as $\mathcal{D}_{QA}$[1]. The label distribution for a sampled set of 10 questions from $\mathcal{D}_{QA}$ is detailed in Appendix A.2.2.

[1]We cannot release the dataset due to proprietary reasons.

3.2 Experimental Setup

We utilize a suite of eight LMs (mix of closed-source and open-source), categorizing them into three groups: *Large*, *Medium*, and *Small*, based on their number of parameters as illustrated in Table 1 (refer Appendix B.2).

Given a question $\mathcal{Q}$ and a conversation $\mathcal{C}$ where $(\mathcal{Q}, \mathcal{C}) \in \mathcal{D}_{QA}$, we prompt a language model, $\mathcal{L}$, to engage in chain-of-thought reasoning (Wei et al., 2022b). The model first identifies evidences relevant to answering $\mathcal{Q}$ based on $\mathcal{C}$, synthesizes these evidences, and finally concludes the answer $\mathcal{A}$ based on the synthesized reasoning. This approach mirrors the annotation guideline provided to annotators, ensuring consistency with human reasoning processes (refer Figure 6 in Appendix B.3 for the prompt template). We hypothesize that this method evaluates the ability of an LM to comprehend contact-center conversations and autonomously reason through them to answer the question $\mathcal{Q}$ based on identified evidences and synthesis. Finally, we report the performance of model $\mathcal{L}$ on $\mathcal{D}_{QA}$ in terms of Macro F1, evaluated over annotated labels in Section 3.1. Refer to Table 1 for detailed results across the suite of eight models.

3.3 Results

The results from Table 1 reveal a strong correlation between the size of LMs and their performance on the QA task within the contact-center domain. We observe that larger models consistently outperform the smaller ones, indicating that they possess more robust domain-knowledge of contact-centers. Specifically in the *Large* group, we note the highest Macro F1 of 75.48% using Claude-3.5-Sonnet (Anthropic, 2023), followed closely by Llama3-70B (Touvron et al., 2023). Notably, GPT-4o (OpenAI, 2023), while being the largest of the lot, performs significantly lower than Claude-3.5-sonnet. We hypothesize that this could be attributed to differences in their training data and methodology.

Interestingly, despite being in the *Medium* group, GPT-4o-mini performs marginally better than GPT-4o. We hypothesise that this could possibly be due to sensitivity to inference parameters, such as prompt template, temperature, maximum target tokens, etc. However, we leave this exploration as a part of future scope and thereby maintain fairness in benchmarking by utilising the same inference parameters across all models.

Additionally, the *Small* group, represented by

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: Illustrates that large LMs generally outperform smaller ones on contact-center QA task indicating a strong correlation between model size and performance, underscoring the proficiency of large LMs in comprehending contact-center conversations from QA standpoint.

Phi-3-mini-128k-instruct (Abdin et al., 2024) and Gemma-2B-it (Mesnard et al., 2024) , has the lowest scores of 62.91% and 54.17%, respectively. Since we do not provide any domain-specific inputs (except the QA question) while inferring using these models, these results highlight the significant performance gap between smaller and larger LMs, suggesting that smaller LMs lack the extensive domain-knowledge inherent in the larger LMs. Consequently, smaller LMs would likely need to rely on external mechanisms, to distill the requisite contact-center-specific knowledge.

4 Distilling Domain-Knowledge To Small LMs

Given that large LMs demonstrate proficiency in contact-center domain-knowledge, we explore the feasibility of transferring this to smaller LMs. Specifically, we select Phi-3-mini-128k-instruct (Abdin et al., 2024) as our target model due to its superior performance among the *Small* group. However, our approach is generic enough to be extended to any LM.

4.1 Experimental Setup

To investigate the effectiveness of transferring contact-center domain-knowledge from large LMs to smaller LMs, we implement the following experimental setups:

4.1.1 Inference With Large LM Guided Plan

In this setup, we follow a two-step process wherein we first utilize a large LM $\mathcal{M}$, proficient in contact-center domain-knowledge to generate an evaluation plan $\mathcal{P}$ in response to a question $\mathcal{Q}$, outlining the criteria for evaluation. We hypothesize, that this

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 2: Example evaluation plan to assess an agent on: Did agent demonstrate active listening ?

plan $\mathcal{P}$ not only provides a structured evaluation criteria for $\mathcal{Q}$ but also breaks it down into simpler components that can be easily comprehended by smaller LMs (refer to Figure 2). Subsequently, given $\mathcal{Q}$, conversation $\mathcal{C}$, and the generated plan $\mathcal{P}$, we then prompt Phi-3-mini-128k-instruct (henceforth, referred to as $\mathcal{M}_{Phi}$) out-of-the-box to engage in chain-of-thought reasoning analogous to that described in Section 3.2. Refer to Figure 7 and Figure 8 in Appendix B.3 for the prompt templates illustrating the generation of evaluation plans and final inference, respectively. Since Claude-3.5-Sonnet (henceforth, refer to as $\mathcal{M}_{Sonnet}$) demonstrates best proficiency in domain-knowledge (refer to Table 1), we fix $\mathcal{M} = \mathcal{M}_{Sonnet}$ for this setup. We hypothesize that the generated plan plays a crucial role in bridging the gap between the domain-knowledge of $\mathcal{M}$ and $\mathcal{M}_{Phi}$, thereby enhancing the ability of $\mathcal{M}_{Phi}$ to reason and answer QA questions effectively.

4.1.2 Fine-tuning With Large LM Generated Response

To further explore the integration of contact-center domain-knowledge, we conduct in-domain fine-tuning of $\mathcal{M}_{Phi}$ on the QA task. Instead of manually annotating a large dataset for fine-tuning, which is resource-intensive, we once again leverage $\mathcal{M}_{Sonnet}$ to generate chain-of-thought reasoning and answer for 780 additional questions across approximately 100 interactions each, following a similar methodology as described in Section 4.1.1 and utilise it as the ground truth for fine-tuning. We randomly sample 80% of questions from this and include all the corresponding examples in training

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 plans generated by a more knowledgeable model $\mathcal{M}_{Sonnet}$ not only enhances smaller models' proficiency in understanding contact-center conversations out-of-the-box, but also plays a crucial role in fine-tuning smaller models on QA task.

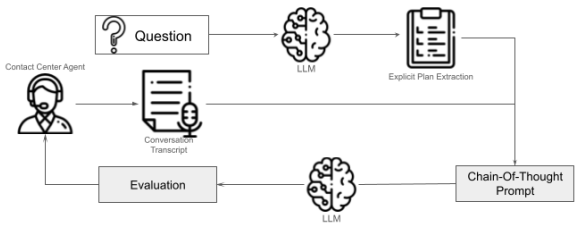

Figure 3: Flow diagram illustrating the two-step QA process: (1) Generating an evaluation plan using a large LM, followed by refinement with human-in-the-loop feedback, and (2) Evaluating the agent based on the refined plan and given conversation.

set (henceforth, referred to as $\mathcal{D}_{Train}$), whereas remainder of the dataset is utilised as the development set $\mathcal{D}_{Dev}$. Subsequently, given a question $\mathcal{Q}$, a conversation $\mathcal{C}$, and a plan $\mathcal{P}$, we perform supervised fine-tuning of $\mathcal{M}_{Phi}$ to generate an output $\mathcal{O}$, where $\mathcal{O}$ aligns with the output generated by $\mathcal{M}_{Sonnet}$. The fine-tuned model is then evaluated on $\mathcal{D}_{QA}$ in terms of Macro F1. Finally, we summarise the results in Table 2.

4.2 Results

For setup S1, we observe that inference using $\mathcal{M}_{Phi}$ guided by evaluation plan generated with $\mathcal{M}_{Sonnet}$ outperforms out-of-the-box inference using $\mathcal{M}_{Phi}$ (setup S0), as illustrated in Section 6, by over 5%. This demonstrates that a simple yet effective idea of chain-of-thought reasoning combined with an evaluation plan from a more knowledgeable model helps in bridging the gap in their domain-knowledge. Additionally, it also highlights that inference using a large LM guided plan can potentially be a promising approach to distill domain-knowledge into smaller LMs, specifically in resource-constrained scenarios where explicit domain-specific fine-tuning is not feasible.

Fine-tuning $\mathcal{M}_{Phi}$ with responses generated by $\mathcal{M}_{Sonnet}$ (S2) yields a substantial improvement of 13.87% over S1 and 18.95% over S0. This indicates that in-domain fine-tuning using silver-data generated by a more knowledgeable LM can effectively transfer domain-knowledge of contact-centers, significantly enhancing the smaller model's performance while eliminating the need for time-consuming gold-standard data collection with human annotations.

Additionally, we also perform an ablation study to understand the importance of individual components in the fine-tuning process. Specifically, we fine-tune $\mathcal{M}_{Phi}$ with evidences and synthesis as target response, while excluding the evaluation plan and note a drop in Macro F1 by approximately 3% (S2 versus S2a). In contrast, fine-tuning with synthesis alone as target response (excluding evidences) along with $\mathcal{M}_{Sonnet}$ generated plan results in only a marginal drop in Macro F1 of 0.28% (S2 versus S2b). This further reinforces the critical role of evaluation plan in guiding the model's reasoning process. While our current benchmarking primarily utilises the final-answer concluded using the chain-of-thought reasoning, evaluation of the generated reasoning (evidence and synthesis) on the grounds of its faithfulness, factual consistency and completeness poses another dimension to study the effectiveness of our approach. However, we leave this exploration as a part of future work and at the same time wish to draw the attention of research communities along this direction.

Moreover, incorporating the evaluation plan into the inference process naturally extends to a human-in-the-loop setting, where the plan can be further refined with human feedback to enhance the domain-specific capabilities of smaller LMs beyond those of large LMs. The flow diagram in Figure 3 illustrates a two-step evaluation process, beginning with generating an evaluation plan using a large language model, followed by refining this plan with human input. Once defined, these evaluation plans can be saved as a one-time process aligned with the defined questions. For every incoming interaction the agent handles, the pre-defined plan can then be utilized for evaluation. This ensures that the assessment is consistent and contextually aware, leveraging the combined strengths of LMs and human expertise for a continual evaluation process.

Finally, while the inclusion of evidence in the generated response has only a marginal impact on

fine-tuning performance, we hypothesize that it significantly aids the interpretability of model responses. This makes it a crucial component for building user trust in the generated model outputs.

5 Prior Work

Language Models (LMs) have shown considerable advancements in recent years, demonstrating their ability to generate fluent text across a wide range of inputs (Wei et al., 2022a; OpenAI, 2023). These advancements have fueled significant interest in applying LMs to domain-specific contexts, where fine-tuning general-purpose models with domain-specific data has led to notable performance improvements across various specialized fields such as legal, medical, and finance domains, highlighting their ability to adapt and perform complex tasks. Notable examples include BioGPT and Med-PaLM in biomedical research (Luo et al., 2022; Singhal et al., 2022), CodeT5 and CodeLLaMa in coding (Wang et al., 2021; Rozière et al., 2023), and Bloomberg-GPT in finance (Wu et al., 2023). Research into the knowledge embedded within LMs has underscored their vast repository of general information, suitable for diverse applications (Petroni et al., 2019; Yu et al., 2023). Studies have also indicated that the ability of LMs to store and effectively use this knowledge scales with their size, enabling them to handle increasingly complex tasks (Wei et al., 2022a; Roberts et al., 2020). Nevertheless, the performance of these models in contact-center environments, particularly in quality assurance (QA), remains relatively unexplored.

Advanced question-answering techniques, including chain-of-thought (Wei et al., 2022b; Kim et al., 2023), tree-of-thought (Yao et al., 2023), and program-of-thought (Chen et al., 2022), have demonstrated potential in enhancing the reasoning ability of LMs. These methods utilize structured reasoning paths to guide models through multi-step problem-solving processes, thereby enhancing the reliability of their responses. However, these techniques have primarily been explored in contexts such as mathematical, symbolic, and commonsense reasoning. Their direct application to leverage the world knowledge embedded in LMs for domain-specific question-answering in contact-centers warrants further investigation.

Over time, enhancing service quality and customer satisfaction have remained focal points of research within the contact-center industry. Researchers are continuously introducing mechanisms to monitor these in real-time and post-call scenarios. For instance, Roy et al., 2016 introduced a real-time quality assurance system employing statistical and rule-based NLP to enable supervisors to monitor ongoing conversations and intervene as needed. Quality assurance practices in contact-centers traditionally include sentiment analysis (Fu et al., 2022), emotion recognition (Girish et al., 2022), and compliance management (Guruju and Vepa, 2021). Moreover, Ingle et al., 2023 proposed fine-tuning a RoBERTa-style language model to analyze silences within contact-center conversations, offering proactive feedback to agents and enhancing their performance. However, the integration of LMs into contact-center workflows holds significant potential to revolutionize the sector.

Recent studies explore various methods to transfer reasoning abilities from large models to smaller ones. For instance, Deng et al., 2023 experiment with implicit reasoning distilled from a teacher model's hidden states, enabling effective task solving without explicit chain-of-thought reasoning. Similarly, Li et al., 2023 introduce Symbolic Chain-of-Thought Distillation (SCoTD), enhancing smaller models' performance by training on rationalizations from larger models. Additionally, Chen et al., 2024 propose a multi-task learning framework to distill chain-of-thought reasoning, optimizing the integration of reasoning capabilities into smaller models for improved performance. These techniques can be particularly beneficial in resource-constrained environments where deploying large LMs may not be feasible.

6 Conclusion

Our study evaluates eight language models (LMs) for contact-center quality assurance, revealing a strong correlation between model size and performance. Claude-3.5-Sonnet, from the *Large* group, demonstrated superior proficiency. We propose methods to distill domain knowledge into smaller models, achieving up to 18.95% improvement in Macro F1. Using evaluation plans generated by more knowledgeable models enhances smaller models' understanding of contact-center conversations. This approach can be further refined through human-in-the-loop feedback, potentially surpassing larger models' capabilities. Our ablation study emphasizes the critical role of evaluation plans in guiding smaller models' reasoning. These

findings suggest promising avenues for developing AI-assisted evaluation tools in contact-centers, potentially leading to more objective, consistent, and scalable assessment processes.

Ethical Considerations

The proposed method for automatic evaluation of agents raises several ethical concerns that must be carefully addressed. We outline these considerations and propose mitigation strategies below:

1. **Bias and Fairness:** The underlying ASR system utilizes acoustic modeling trained on US-English dialects. To mitigate potential biases:
    - We do not recommend using this system for non-US-English conversations.
    - For adaptation to other dialects or languages, developers must ensure careful curation of training data and adopt strategies to eliminate biases towards particular groups.
    - Regular audits should be conducted to identify and address any emerging biases in the system.
2. **Human Oversight and Accountability:** Given the impact on employee performance evaluation, compensation, and career growth:
    - Implement a 'human-in-the-loop' mechanism for constant monitoring and intervention.
    - Establish a clear dispute resolution process for employees to challenge machine-generated predictions.
    - QA supervisors should have discretion to utilize or discard model predictions.
    - Regular training for supervisors on the system's capabilities and limitations is essential.
3. **Privacy and Data Security:**
    - Sensitive data is redacted before analysis, ensuring individuals cannot be traced.
    - Implement robust data encryption and access control measures.
    - Regularly audit data handling processes to ensure compliance with privacy regulations.
4. **Transparency and Explainability:**
    - Develop clear communication materials explaining how the system works and impacts evaluations.
    - Provide agents with access to their evaluation data and the factors influencing their scores.
    - Regularly update documentation as the system evolves.
5. **Continuous Improvement:**
    - Establish a feedback loop to continuously improve the system's accuracy and fairness.
    - Regularly update the system to address identified biases, errors, or new ethical concerns.

By implementing these ethical considerations, we aim to create a more fair, transparent, and accountable automated evaluation system that respects employee rights and privacy while providing valuable insights for quality assurance. It is crucial to continually reassess and adapt these considerations as the technology and its applications evolve.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

pages 20–24, Melbourne, Australia. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Anthropic. 2023. Model Card: Claude 3 technical report. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. [Online; accessed 18-July-2024].

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *CoRR*, abs/2211.12588.

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024. Learning to maximize mutual information for chain-of-thought distillation. *CoRR*, abs/2403.03348.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart M. Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *CoRR*, abs/2311.01460.

Xue-Yong Fu, Cheng Chen, Md. Tahmid Rahman Laskar, Shayna Gardiner, Pooja Hiranandani, and Shashi Bhushan TN. 2022. Entity-level sentiment analysis in contact center telephone conversations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, pages 484–491. Association for Computational Linguistics.

K. V. Vijay Girish, Srikanth Konjeti, and Jithendra Vepa. 2022. Interpretabilty of speech emotion recognition modelled using self-supervised speech and text pre-trained embeddings. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4496–4500. ISCA.

Sai Guruju and Jithendra Vepa. 2021. Addressing compliance in call centers with entity extraction. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2347–2348. ISCA.

Digvijay Ingle, Ayush Kumar, and Jithendra Vepa. 2023. Listening to silences in contact center conversations using textual cues. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 2688–2692. ISCA.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12685–12708. Association for Computational Linguistics.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2665–2679. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6).

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Varun Nathan, Ayush Kumar, and Digvijay Ingle. 2024. Can probing classifiers reveal the learning by contact center large language models?: No, it doesn't! In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 92–100, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Shourya Roy, Ragunathan Mariappan, Sandipan Dandapat, Saurabh Srivastava, Sainyam Galhotra, and Balaji Peddamuthu. 2016. Qa^{rt}: A system for real-time holistic quality assurance for contact center dialogues. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3768–3775. AAAI Press.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Yue Wang, Weishi Wang, Shafiq R. Joty, and Steven C. H. Hoi. 2021. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8696–8708. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023. Kola: Carefully benchmarking world knowledge of large language models. *CoRR*, abs/2306.09296.

Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zifan Chen, Yi Xie, Jie Zhao, Quanzheng Li, Yang Chen, Li Zhang, Lin Shen, and Bin Dong. 2024. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Medicine Plus*, 1(2):100030.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *Preprint*, arXiv:2305.15005.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A Data Annotation

A.1 Annotation Samples

In Figure 4 and Figure 5, we present example pairs of a transcript and a question with corresponding annotated answers.

```
Transcript:
customer: hello
agent: hello this is *** from my *** on a recorded
line is this ***
customer: yes
agent: hi *** we're the company that handled the
warranty program on your hearing aids and ...
customer: i don't know but i i didn't get what you
said who who are you with now
agent: my essentials and we're the company that
carry the warranty program on your hearing aids you
have the secure plan with us
customer: i still didn't get
agent: okay what i can do is reach out to your
hearing aid provider for them to reach out to you
okay
customer: yeah would it send me something that i
could read
agent: okay sir i will do that
customer: alright yeah
agent: alright bye bye

Question:
does the agent overcome all the customers
questions and objections?

To answer the given question, let's think step by
step:
Evidences:
1. The customer raised the following questions/
objections during the conversation: "i don't know
but i i didn't get what you said who who are you
with now", "i still didn't get"
For the first question, the agent provided an
explanation: "my essentials and we're the company
...". However, the customer still did not understand
, indicating that the response was not satisfactory.
Synthesis:
Based on the evidences, the agent was able to
identify the customer's questions and objections.
However, the agent's responses did not fully address
 or overcome those questions and objections in a
clear and satisfactory manner.
Hence, the final answer is: no

Answer: no
```

Figure 4: Sample abridged transcript with annotated label as No

```
agent: hi this is *** my essentials on a recorded
line is this ***
customer: yes
agent: hi *** we are the company that handles your
warranty program on your hearing aids and ...
customer: okay now can you miss the appointment
agent: i can transfer you to *** if you would like
customer: okay
agent: alright bear with me one moment
customer: yes
agent: i did try to reach out to him i got their
voice mail but i left a voice mail ask them to reach
 out to you okay
customer: okay do you
agent: alright
customer: have the phone number please
agent: yes let me just find that for you real quick
customer: mh
agent: oh let's see if i can spell right today it
would be great oh my goodness alright their phone
number is ***
customer: thank you very much
agent: you're welcome you have a wonderful day

Question:
does the agent overcome all the customers
questions and objections?

To answer the given question, let's think step by
step:
Evidences:
1. The customer asked "okay now can you miss the
appointment" which can be interpreted as a question
about making an appointment. The agent responded by
offering to transfer the customer to the hearing aid
 provider's office to make the appointment: "i can
transfer you to *** if you would like".
2. After being unable to connect the call, the agent
 left a voicemail with the customer's details and
the request to make an appointment for the 10-point
inspection and discuss further coverage: "she needs
to make an appointment because her...".
Synthesis:
The agent identified and acknowledged the customer's
 main question/objection about making an appointment
 for the 10-point inspection and discussing further
coverage. The agent attempted to address this by
offering to transfer the call to the hearing aid
provider's office, and when that failed, left a
detailed voicemail with the customer's information
and the request to make the appointment.
Hence, the final answer is: yes

Answer: yes
```

Figure 5: Sample abridged transcript with annotated label as Yes

A.2 Dataset Statistics

In this section, we report additional dataset statistics of the transcript length and label distribution of curated datasets used for evaluating and fine-tuning models.

A.2.1 Transcript length Distribution

Table 3 shows the distribution of transcript length in terms of number of words with the mean, standard deviation, min-max, and interquartile range statistics.

Metric	$\mathbf{D_{QA}}$	$\mathbf{D_{Train}}$
25%	495.00	536.00
50%	815.00	873.00
75%	1280.00	1493.00
mean	1026.62	1420.00
std	814.50	1774.97
min	8.00	2.00
max	4885.00	41484.00

Table 3: Transcript length statistics.

Question	No	Yes	Total
did the agent align with customer on reason for call and assure them that they will be able to assist them or that they will get them on the line with the best person to assist them?	22	54	76
did the agent accurately provide next payment date and amount?	25	75	100
did the agent follow the correct process/procedure for a new customer?	34	62	96
was the agent able to refrain from disclosing the customer's phone number in the database?	4	62	62
did the agent properly acknowledge customer inquiry?	2	98	100
did the agent attempt to verify the customer's contact information?	13	49	62
did the agent offer an approved assuring statement?	98	2	100
did the agent clearly explain deposit/cancellation policies as listed under property policies and fees?	82	18	100
did the agent avoid interrupting or talking over customer and show active listening skills?	2	97	99
did the agent ask the customer to take a moment for a brief survey after the call?	71	29	100

Table 4: Label distribution of 10 sampled questions from the training dataset.

A.2.2 Label Distribution

$\mathcal{D}_{Train}$ has a balanced distribution of target labels with 51.36% of `yes` and 48.64% of `no` labels. Such balance ensures that the fine-tuned model is not likely to be biased toward predicting any one class. $\mathcal{D}_{QA}$ has 56.15% of `yes` and 43.85% of `no` labels. We report the distribution of the labels of a sample of 10 questions $\mathcal{D}_{QA}$ in Table 4.

Model	Price ($) per 1M Tokens	
	Output	Input
GPT-4o	15	5
Claude-3.5-Sonnet	15	3
GPT-4o-mini	0.6	0.15
Llama3-70B	3.5	2.65
Llama3-8B	0.6	0.3
Mistral-7B	0.2	0.15

Table 5: Pricing for API usage.

B Model Inference Details

B.1 Inference parameters

We use the OpenAI and Amazon Bedrock APIs to run inference for the Large and Medium LMs described in Section 3.2. To infer with the Small LMs, i.e., Phi-3 and Gemma models, we host the LMs on an AWS EC2 instance with an NVIDIA Tesla A100 GPU having 80GB GPU memory. We set *max_new_tokens* to 1024 and *temperature* to 0 for all models.

B.2 API Usage Pricing

For *GPT-4o*, *GPT-4o-mini*, and *Claude-3.5-Sonnet*, we do not have visibility into their number of parameters, hence, we use their respective pricing for API usage via OpenAI[2] and Amazon Bedrock[3] as of July 18, 2024 as a proxy to assign the appropriate group in Table 1. We tabulate the pricing in Table 5 for reference.

B.3 Prompt Templates

In this section, we provide various prompts used in the experiments. The prompt template for implicit CoT reasoning discussed in Section 3.2 is presented in Figure 6. The prompt template for evaluation plan generation and inference with Large LM guided plan discussed in Section 4.1.1 is presented in Figure 7 and Figure 8, respectively.

```
As a call center QA expert, evaluate an agent's
interaction based on:

1. Given question
2. Conversation transcript
3. Answer options

Analyze the conversation and provide a step-by-step
response:

1. Evidences: List relevant points from the
conversation
2. Synthesis: Summarize your rationale
3. Conclusion: State the final answer

Format your response as follows:

To answer the given question, let's think step by
step:

Evidences:
- Evidence 1
- Evidence 2
...

Synthesis:
(Summarize your reasoning)

Hence, the final answer is: (Your chosen answer)
```

Figure 6: Implicit CoT Reasoning prompt template.

[2]https://openai.com/api/pricing/

[3]https://aws.amazon.com/bedrock/pricing/

```
As a call center QA expert, break down the given
evaluation question into criteria for assessing
agent performance. Criteria should be:
- Determinable from the conversation alone
- Unique and non-repetitive
- Clear and concise

Provide a Python-parsable JSON response in this
format:

[
  {
    "name": "<criterion_name>",
    "description": "<criterion_description>",
  },
  ...
]

Include only the JSON object in your response.
```

Figure 7: Plan Generation prompt template.

```
As a call center QA expert, evaluate an agent's
interaction based on:

1. Main question
2. Sub-criteria
3. Conversation transcript
4. Answer options

Analyze the conversation and provide a step-by-step
response:

1. Evidences: List relevant points for each sub-
criterion
2. Synthesis: Summarize your rationale
3. Conclusion: State the final answer

Format your response as follows:

To answer the given question, let's think step by
step:

Evidences:
(List evidences for each sub-criterion)

Synthesis:
(Summarize your reasoning)

Hence, the final answer is: (Your chosen answer)
```

Figure 8: CoT Reasoning with Plan prompt template.

C Fine-Tuning Details

C.1 Prompt Templates

Given a question $\mathcal{Q}$, a conversation $\mathcal{C}$, and a plan $\mathcal{P}$, $\mathcal{M}_{Phi}$ is fine-tuned to generate an output $\mathcal{O}$ containing answer and associated reasoning (evidence and synthesis). We followed similar prompt templates as described in Section B.3 to generate the plan and reasoning.

C.2 Hyperparameters and Infrastructure

In order to fine-tune $\mathcal{M}_{Phi}$ model for the QA task, we utilise the Phi-3-mini-128k-instruct[4] checkpoint from the HuggingFace library (Wolf et al., 2019).

[4]https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

The fine-tuning process is carried out on a single NVIDIA A100 80GB GPU, employing the training dataset ($\mathcal{D}_{Train}$) curated as detailed in Section 4.1.2. To identify the optimal hyperparameters, we perform a grid search across several configurations. The hyperparameter space included: learning rate $\in \{1e-6, 5e-5, 1e-5\}$, batch size $\in \{4, 8\}$, a fixed number of epochs set to 2, and a warmup ratio of 0.05. We choose, the best model checkpoint based on evaluation loss computed on the validation set ($\mathcal{D}_{Dev}$). Finally, we choose the model configuration yielding highest Macro F1 score on $\mathcal{D}_{QA}$ for final evaluation, ensuring optimal performance for the contact-center evaluation task.

D Domain Knowledge in Contact-Center QA

We refer to domain knowledge in Contact-Center QA in two distinct ways:

- **Industry-Specific Knowledge:** This refers to an understanding of information that pertains to a particular industry or sector, such as general concepts, terminology, and practices common to that domain. For instance, in the banking sector, this could include knowledge about general banking operations, financial terms, or customer service practices. Larger models, such as Claude-3.5-Sonnet, often perform better in this area due to their broad pretraining on diverse datasets that encompass general industry-specific contexts.
- **Conversational Language Understanding:** This aspect of domain knowledge involves the ability to comprehend and interpret conversational language used between agents and customers, which may include resolving misunderstandings, addressing customer concerns, or adapting to various tones and styles of communication. While this type of knowledge is not tied to specific products or services, it is equally crucial in the evaluation of contact-center interactions, as it helps assess how well an agent navigates the conversation.

In our experiments, both types of knowledge are essential for evaluating agent performance in contact centers, and larger models often demonstrate more robust comprehension in these areas. By transferring such knowledge from larger models to smaller models through evaluation plans, we

aim to enhance the latter's ability to perform both product/service-specific reasoning and conversational language understanding.