# Using Advanced LLMs to Enhance Smaller LLMs: An Interpretable Knowledge Distillation Approach

Tong Wang, K. Sudhir and Dat Hong\*
Yale School of Management
tong.wang.tw687@yale.edu, k.sudhir@yale.edu, dat.hong@yale.edu

Advanced Large language models (LLMs) like GPT-4 or LlaMa 3 provide superior performance in complex human-like interactions. But they are costly, or too large for edge devices such as smartphones and harder to self-host, leading to security and privacy concerns. This paper introduces a novel interpretable knowledge distillation approach to enhance the performance of smaller, more economical LLMs that firms can self-host. We study this problem in the context of building a customer service agent aimed at achieving high customer satisfaction through goal-oriented dialogues. Unlike traditional knowledge distillation, where the "student" model learns directly from the "teacher" model's responses via fine-tuning, our interpretable "strategy" teaching approach involves the teacher providing strategies to improve the student's performance in various scenarios. This method alternates between a "scenario generation" step and a "strategies for improvement" step, creating a customized library of scenarios and optimized strategies for automated prompting. The method requires only black-box access to both student and teacher models; hence it can be used without manipulating model parameters. In our customer service application, the method improves performance, and the learned strategies are transferable to other LLMs and scenarios beyond the training set. The method's interpretabilty helps safeguard against potential harms through human audit.

Key words: Large Language Models, Knowledge Distillation, Interpretability, Customer Satisfaction

<sup>\*</sup> The authors are grateful for insightful comments from participants at the 2024 Biz AI conference, 2024 Four Schools (Columbia, NYU, Wharton, Yale) Conference, 2024 HongKong Quant Marketing Mini-Conference, Summer Workshop on AI for Business 2024, 2024 ISMS Doctoral Symposium in Sydney, 2024 Monash Marketing Mini-Symposium, Yale Quantitative Marketing Brown Bag Seminar, marketing seminars at HKU and CUHK.

### 1. Introduction

Advancements in large language models (LLMs) have enabled low-cost automation of many marketing tasks. Choosing the right LLM for a task involves balancing cost, feasibility, and performance. While high-performance models like GPT-4 are appealing and may seem like the obvious choice, many businesses prefer smaller, more affordable, or open-source LLMs. In high volume applications such as customer service, even small cost differences per query can lead to large differences in total cost; as such cheaper or free models like LlaMa may offer a better balance in tradeoffs, despite lower performance. Additionally, data privacy concerns drive firms to self-host LLMs instead of using third-party APIs, which could compromise data confidentiality. Consequently, there is growing interest in enhancing the performance of smaller, cost-effective LLMs.

This paper examines whether we can augment the effectiveness of smaller, but more economical LLMs using the knowledge embedded in larger and superior LLMs. This concept is called knowledge distillation, where a more advanced LLM, referred to as the "teacher," imparts its knowledge to a less sophisticated LLM, termed the "student," to enhance its performance.

While enhancing a weaker LLM through a more advanced LLM is a general problem applicable in many contexts, in this paper, we illustrate it for the problem of goal-oriented dialogue tasks (e.g. Wei et al. 2018, Bordes et al. 2016). Goal-oriented dialogues aim to steer conversations toward specific desired outcomes (Ham et al. 2020a); as such it is relevant for conversations involving negotiations (Samad et al. 2022) and persuasion (Wang et al. 2019). Specifically, we consider a customer service application, where an LLM acts as a customer service agent that interacts with the customer with the goal of achieving high customer satisfaction.

A common strategy for improving student performance through knowledge transfer from a teacher is fine-tuning, where the teacher generates training data to train the student (e.g., Tang et al. 2019, Agarwal et al. 2023, 2024). However, this method has significant limitations. Firstly, fine-tuning requires access to and updates of model parameters, which is not always feasible, especially for LLMs that only allow API access; and also costly since it involves updating billions of parameters. Secondly, the distilled "knowledge" is encapsulated in model parameters, making it unintelligible to humans. This opacity hinders debugging and raises safety concerns, particularly when the teacher is an external, unverifiable model. The inaccessibility of distilled knowledge complicates maintenance, as tracking changes and ensuring consistency with previous iterations can be problematic. Finally, fine-tuning focuses only on lexical similarities in the outputs, ignoring the underlying strategies that lead a model to produce specific responses and the fact that different lexical choices can express identical meanings or follow the same response strategy.

A student model can also be improved via prompt tuning (Lester et al. 2021, Hu et al. 2021), also known as soft prompt tuning), where specific prompt tokens are fine-tuned to elicit better performance. This involves adjusting the prompt to align with desired outputs, enhancing the model's ability to generate accurate and contextually appropriate responses. However, prompt tuning lacks interpretability, as the prompts are represented as embeddings that are hard to understand. Additionally, a single, fixed prompt tuned for an entire task cannot capture the richness and diversity of strategies needed for different scenarios. Furthermore, prompt tuning requires access to the LLM's internals, making it inapplicable to LLMs that are only accessible via APIs.

Therefore, we propose an interpretable knowledge distillation method. Instead of directly distilling teacher's "knowledge" into student's model parameters, which would require

accessing and updating the student model, our method constructs a knowledge base that the student can query externally without altering the student model itself. This knowledge base, which we refer to as a *library*, consists of representative *scenarios* that the student could encounter during deployment, along with corresponding strategies. Each scenario is represented by an on-going dialogue between an agent and a customer, and the corresponding strategies guide the student on *how* to respond when continuing the dialogue. Then, during deployment, the library functions similarly to retrieval-augmented generation (RAG) (Lewis et al. 2020); the student LLM identifies the most relevant scenarios through embeddings and applies the corresponding strategies to stimulate an appropriate response.

Building a library for a multi-step, goal-oriented dialog problem like customer service presents additional challenges. First, generating scenarios involves interactive dialogues between the teacher and a customer to capture diverse conversational paths. However, during deployment, the student cannot perfectly replicate the teacher's quality of interactions, leading to deviations. In multi-step settings such as conversations, even minor quality differences from the teacher's responses at a given stage can lead to more negative customer response, and this negativity can cumulatively amplify over the entire conversation. As such, a student-customer conversation can veer into unencountered scenarios if the library is constructed based on only teacher-customer conversation scenarios. This phenomenon, known as distribution shift, poses a significant challenge when the student's training data is generated exclusively from the teacher. Second, the strategies in the library must be adaptable to each student and specific scenario. These strategies should guide the student's behavior by considering the student's current capabilities and providing targeted suggestions. Finally, a third important issue is that agents must adhere to firm policies, business requirements, and cost constraints when addressing customer requests. Thus, these policies and constraints must be integrated into the strategy development process to ensure the agent's solutions are practical and compliant with firm policies.

We design a novel method to address the aforementioned challenges. Our method is an iterative process where, in each iteration, a new batch of scenarios and their corresponding strategies are added to the library. To generate scenarios, both the student and teacher interact with the environment (e.g., a donor in a persuasion task or a customer in customer service) to produce conversations and sample scenarios. To address the issue of distribution shift, we ensure that enough interactions involve the student by gradually increasing the probability of selecting the student for scenario generation, so that the student eventually dominates this process. In the strategy learning phase, strategies are generated and refined iteratively. The teacher LLM evaluates both its own and the student's responses, providing targeted strategies for the student to follow. These strategies are incorporated into prompts for subsequent refinement rounds, progressively honing the student's ability to mimic the teacher. The library grows iteratively, with the teacher monitoring the student's progress and deciding when to terminate the process based on the need for further feedback or significant improvement. The output is a customized library of scenarios and strategies, containing context-specific knowledge from the teacher that is most relevant to the student and the task.

Our interpretable knowledge distillation approach offers several advantages over finetuning. First, by teaching strategies rather than responses, our method enables LLMs to understand how to handle different scenarios at a strategic level, providing a global view of the problem rather than simply mimicking lexical choices. Second, our method ensures easy transferability across LLMs and contexts. The strategy library allows LLMs to adapt to new, unseen situations, broadening their problem-solving capabilities. This library can be used by other student LLMs or for related customer service problems, potentially enhancing performance without direct training the student LLMs or on specific tasks. Third, the interpretability of our approach significantly enhances AI safety. By extracting explicit strategies, domain experts can review and understand the LLMs' decision-making processes. This transparency facilitates trust and provides safeguards against misuse, errors, or adversarial influences. Finally, as our method does not require access to or modification of the student or teacher LLMs' parameters (both can be queried solely via black-box access, such as APIs), it distinguishes itself from traditional knowledge distillation methods such as prompt tuning and fine-tuning, which require internal model access or parameter updates. This makes it particularly suitable for environments where direct model manipulation is impractical or restricted.

Based on our empirical application, i.e., multi-turn conversations in customer service, our key findings are as follows: First, teaching strategy is more effective than teaching responses for multi-turn generation. Second, context-specific strategies are more effective than global strategies, since the former can provide more targeted strategies for different scenarios. Third, even though the library is learned for a particular student LLM and specific contexts, it contains common knowledge that is transferrable across models and across contexts.

The rest of this paper is organized as follows: §2 discusses the related literature on LLM and knowledge distillation. §3 presents the proposed method, while §4 evaluates it with an extensive set of experiments. §5 concludes the paper. More analyses are included in the online appendix.

### 2. Related Work

Our work contributes to the recent but growing literature on marketing applications using LLMs. This body of research has typically focused on how LLMs can be used for market

research and the study of human behavior, while also highlighting attendant challenges (Gui and Toubia 2023, Qiu et al. 2023, Horton 2023)). Specific marketing research applications include perceptual maps (e.g., Li et al. 2024) and conjoint analysis (e.g., Brand et al. 2023, Gui and Toubia 2023). In contrast, this paper is focused on how to effectively adapt and engineer an LLM for marketing tasks such as customer service. We will next discuss how our method relates to existing literature on knowledge distillation in LLMs and advances the literature on goal-oriented dialogs, particularly in multi-turn interactions.

# 2.1. Knowledge Distillation for LLM

The concept of utilizing a superior model to enhance a less powerful one is known as knowledge distillation. This technique was first introduced by Hinton et al. (2015) in the context of supervised learning and has been adapted for use with language models in recent years (Sanh et al. 2019, Sun et al. 2019). In the realm of language models, knowledge distillation involves using a larger and more capable LLM to generate data that trains specialized, smaller models. Existing research on knowledge distillation for language models typically employs objective functions that either maximize the likelihood of highprobability sequences generated by the teacher model (Kim and Rush 2016) or guide the student model to mimic the token-level probability distributions provided by the teacher (Sanh et al. 2019). Some recent work also proposes to teach the student the rationale for solving a task (Hsieh et al. 2023, Magister et al. 2022). However, all these methods require training the student model and updating its parameters. In our approach, however, we distill the knowledge into an external library that the student can query during inference, without the need for training the student model. The teacher's knowledge is utilized by the student through retrieval-augmented generation (RAG), another popular technique in LLM. In doing so, our method requires only black-box access to the student model, such as through an API, which is not feasible with existing knowledge distillation techniques.

# 2.2. Goal-Oriented Dialogues

Recent advancements in LLMs have significantly improved their application in complex goal-oriented dialogues (e.g., Ham et al. 2020b, Li et al. 2023, Snell et al. 2022), but challenges and limitations remain. First, smaller LLMs often lack a strategic understanding of overall dialogue progression and fail to achieve dialogue objectives through multi-turn interactions (Cheng et al. 2024, Deng et al. 2023). Second, the multi-step nature of goal-oriented dialogues fundamentally differs from one-step tasks like text classification and summarization as fine-tuning at each utterance level overlooks the interdependence of multi-turn utterances and the high-level strategy. Zhang et al. (2023) proposes an "Ask an Expert" solution, where a lesser model seeks advice from a better LLM for generating utterances, but this increases inference (i.e., response) time. Finally, some approaches rely on dialogues with specifically annotated strategies for training (e.g., Zhang et al. 2022, 2023, Joshi et al. 2021), but the dependence on labeled datasets creates a significant barrier for practical application.

# 3. Library-based Interpretable Knowledge Distillation

Given a student LLM, denoted as  $S(\cdot)$  (e.g., LlaMa 2 or GPT-3.5), our method involves creating a library consisting of a set of representative scenarios, paired with corresponding strategies constructed by a teacher LLM  $\mathcal{T}(\cdot)$  optimized for instructing the students on *how* to respond in those scenarios. We first set up the learning environment and then describe the algorithm. Then, we will show how the library is used during deployment and explain the benefits of our method.

# 3.1. Learning Environment

We set up a learning environment where the student attempts to improve its performance by mimicking the teacher. For each input, the teacher compares the student's output with its own output to help the student respond like itself. To facilitate effective learning, we simulate a customer using GPT-4, which we denote as  $C(\cdot)$ . The learning of the student relies on the interaction with the customer LLM. This design is motivated by recent research that advocates using LLMs to simulate human responses to reduce costs and improve efficiency (Li et al. 2024), compared to conducting real human studies.

**3.1.1.** Simulating Customers: To simulate the customer, we describe the task in the prompt where we request the LLM to role-play as a customer calling an airline company to request customer service. Here, we focus on a specific context where the customer bought a restricted ticket (non-changeable and non-refundable) and requests to cancel it without penalty.

To increase the heterogeneity of the customers and the richness of the conversations, we vary the customer's **social styles**, **initial emotions**, and **difficulty**. For social styles, we use four types based on the classification from Merrill and Reid (1981). (i) *Driver*: results-driven, confident, and assertive; (ii) *Analytical*: detail-oriented, systematic, and logical; (iii) *Amiable*: cooperative, empathetic, and relationship-focused; and (iv) *Expressive*: enthusiastic, creative, and spontaneous<sup>1</sup>. For initial emotions, we set four different customer emotions when initiating the call: calm, confused, concerned, and frustrated, leading to varied dialogue developments. Additionally, we vary the difficulty level of the customer by including/not including the keyword "demanding" in the customer role description to the LLM. We observe that including "demanding" changes customer behavior, making customers more persistent with their requests. We use q(s, e, d) to represent the prompt for the customer LLM, parameterized by the social style s, initial emotion e, and difficulty level d.

<sup>&</sup>lt;sup>1</sup> We provide more detailed description of each type in the prompt for simulating customers with GPT-4.

Given the four social styles, four initial emotions, and two difficulty levels, we create a total of 32 types of customers. When simulating customers, each of the 32 types generates many different conversations because we set the temperature of the LLM to non-zero.

3.1.2. Simulating Teacher: We then describe how to prompt a teacher LLM to act as an agent. Here we choose the state-of-the-art LLM, GPT-4, as the teacher. We define a base prompt instructing GPT-4 to role-play as a customer service agent. We denote the prompt as  $p_{\text{base}}$ , which contains three key components: role, goal, and constraints.

The role specifies the position or function that GPT-4 needs to assume. For example, prompting GPT-4 to "role-play as a customer service agent" sets the context for the interaction, guiding the model to generate responses suitable for customer service scenarios. The goal establishes the desired outcome and ultimate objective of the interaction, guiding GPT-4 to adjust its strategies accordingly. For instance, if the goal is to achieve high customer satisfaction, GPT-4 will tailor its responses to be more empathetic, helpful, and solution-focused. Including the goal in the prompt is essential because it provides clear direction for the LLM, ensuring that its responses align with the intended results. The constraint establishes the limitations and boundaries that the LLM needs to follow, and as explained earlier, it helps to ensure that the LLM accounts for business constraints and firm policy. As such, the agent is instructed to adhere to these constraints while aiming to achieve high customer satisfaction. This ensures that all agents follow the same company policy, differing only in their communication strategies, which the teacher aims to teach the student.

The base prompt is fixed for the entire task and included in the prompt for both the teacher and student.

# 3.2. Knowledge Distillation

Our approach iterates over three steps: scenario generation, strategy teaching, and goal evaluation, progressively building up a customized library consisting of scenarios and strategies for handling them. We call each execution of the three steps as one iteration of the algorithm. We use  $\mathcal{L}^{(t)} = \{(\mathbf{s}_i^{(t)}, p_{\text{strategy}}(\mathbf{s}_i^{(t)}))_{i=1}^{n_t}\}$  to represent the library at the end of iteration t, where  $\mathbf{s}_i^{(t)}$  represents a scenario indexed by i,  $p_{\text{strategy}}(\mathbf{s}_i^{(t)})$  is the corresponding prompt, and  $n_t$  is the total number of scenarios in the library at iteration t. As t increases, the library grows larger. Below, we detail each step.

**3.2.1. Scenario Generation:** In this step, the LLM agent interacts with the customer  $\mathcal{C}(\cdot)$  to generate conversations, denoted as  $\mathbf{x}^{(t,l)}$ , where t is the iteration index and l is the conversation index.  $\mathbf{x}^{(t,l)}$  consists of a sequence of utterances from the customer  $\mathcal{C}(\cdot)$  and the customer service agent LLM (e.g.,  $\mathcal{T}$  or  $\mathcal{S}$ ):

$$\mathbf{x}^{(t,l)} = (\mathbf{a}_1^{(t,l)}, \mathbf{c}_1^{(t,l)}, \mathbf{a}_2^{(t,l)} \cdots), \tag{1}$$

where  $\mathbf{a}_{k}^{(t,l)}$  represents the agent's utterance at the k-th turn and  $\mathbf{c}_{k}^{(t,l)}$  represents the customer's utterance at the k-th turn. Without loss of generality, we assume a conversation always starts with the agent's utterance, for example, "Hello, how may I help you?"

Since the generation of an utterance depends on all previous exchanges, we define a subconversation with the first k turns, ending with the customer's utterance.

$$\mathbf{x}_{[:k]}^{(t,l)} = (\mathbf{a}_1^{(t,l)}, \mathbf{c}_1^{(t,l)}, \cdots, \mathbf{a}_k^{(t,l)}, \mathbf{c}_k^{(t,l)}). \tag{2}$$

The generation of the teacher's utterance at the k-th turn is based on the prior conversation  $x_{[:k]}^{(t,l)}$  as well as the base prompt  $p_{\text{base}}$ :

$$\mathbf{a}_k^{(t,l)} = \mathcal{T}(\mathbf{x}_{[:k-1]}^{(t,l)}|p_{\text{base}}). \tag{3}$$

The customer's utterance is determined by the prior conversation and the parameterized prompt q(s, e, d):

$$\mathbf{c}_{k}^{(t,l)} = \mathcal{C}(\mathbf{x}_{[:k-1]}^{(t,l)}, \mathbf{a}_{k}^{(t,l)} | q(s, e, d)). \tag{4}$$

For each iteration t, we generate a set of conversations, denoted as  $\mathbf{X}^{(t)}$ , by varying the customer prompt parameters s(social styles), e(initial emotions) and d(difficulty) and running multiple times for each set of parameters with a non-zero temperature.

Subconversations, defined in Equation (2), are randomly sampled from each conversation in  $\mathbf{X}^{(t)}$ . We call these subconversations **scenarios**. Each scenario includes the entire prior conversation ending with the customer's utterance. We use  $\mathbf{S}^{(t)}$  to represent the set of scenarios sampled from  $\mathbf{X}^{(t)}$ .

In a naive solution, the conversations in  $\mathbf{X}^{(t)}$  are generated by  $\mathcal{T}$  interacting with  $\mathcal{C}$ . However, each scenario extracted from  $\mathbf{X}^{(t)}$  reflects the specific dynamics and decisions made by the teacher  $\mathcal{T}$ , not necessarily those a student would encounter or make when interacting independently in similar contexts. This leads to compounding errors because the scenarios the student encounters during deployment will differ from those it observed during training. In a multi-step decision-making process, even a small deviation in one step can compound over subsequent steps, leading to significant differences at the conversation level. This mismatch between training and deployment scenarios is a significant for challenge multi-step imitations, often referred to as distribution shift (Pomerleau 1991, Ross and Bagnell 2010).

To overcome this challenge, we let the student participate in the scenario generation step. We assign a probability p for the teacher to be selected to interact with the customer for conversation generation and 1-p for the student to be selected. Initially, p is set to 1 in iteration 1 and gradually decreases in subsequent iterations.

In each iteration, when the student is selected to interact with the customer LLM, it will produce an output based on the scenarios and strategies retrieved from the current library:

$$\mathbf{a}_k^{(t,l)} = \mathcal{S}(\mathbf{x}_{[:k-1]}^{(t,l)}|p_{\text{base}}, \mathcal{L}^{(t)}). \tag{5}$$

We will explain how to retrieve the relevant strategy from the library in Section 3.3.

As the iterations progress, more scenarios will be accumulated into  $\mathcal{L}^{(t)}$ , covering a more diverse set of situations the student could encounter.

**3.2.2.** Strategy Teaching: In this step, the teacher iteratively generates and refines strategies for each scenario until the student accurately mimics the teacher's output for that scenario. See Figure 1 for an illustration of the process. We call the prompt generated

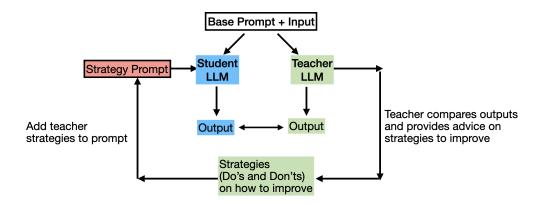


Figure 1 An iterative process of strategy teaching

at this step the *strategy prompt*. The strategy prompt will be combined with the base prompt to instruct the student *how* to behave like the teacher.

Specifically, for each scenario, denoted as  $\mathbf{s}$ , both  $\mathcal{T}$  and  $\mathcal{S}$  generate a response. These responses are then evaluated by  $\mathcal{T}$ , which identifies discrepancies and proposes updates to the strategy. The updates are suggestions made to the student instructing what they should or should not do, based on their current output. They can vary from general, such

as "use a more empathetic tone" to highly nuanced, such as "address the customer by their first name", tailored to the specific demands of each scenario and the difference between the student and the teacher.

These updates are incorporated into the strategy prompt, denoted as  $p_{\text{strategy}}(\mathbf{s})$ .

teacher's output: 
$$\mathbf{a} = \mathcal{T}(\mathbf{s}|p_{\text{base}}).$$
 (6)

student's output: 
$$\mathbf{a} = \mathcal{S}(\mathbf{s}|p_{\text{base}}, p_{\text{strategy}}(\mathbf{s})).$$
 (7)

This process of iterative refinement of  $p_{\text{strategy}}(\mathbf{s})$  continues until there are no further updates from the teacher or a pre-defined maximum number of refinements is reached. By progressively refining its responses based on learned strategies and feedback, the student model becomes more adept at handling the scenarios, aligning its behavior with the teacher's expertise and improving overall performance<sup>2</sup>.

3.2.3. Termination Mechanism: Our main algorithm is an iterative process that gradually grows the library by adding a new batch of scenarios and their corresponding strategy prompts at each iteration. The thorough coverage of the data space by the scenarios is crucial for maintaining the robustness and reliability of the strategy prompts during deployment. To determine when the sufficiency is reached, we incorporate a goal evaluation step in the iterative process, where the student LLM is evaluated, based on the interactions with the customer LLM, after each new batch is added to the library, to determine whether the student LLM has achieved satisfactory performance using the current library.

We choose to use LLMs for the evaluation, motivated by recent research showing the potential of using LLMs to substitute human evaluations in various tasks, particularly in assessing other LLMs, with findings of up to 80% agreement with human judgment.

<sup>&</sup>lt;sup>2</sup> A detailed example of updating the strategies and improving the response is shown in Appendix C.

Leveraging this advancement, our approach employs the teacher, a state-of-the-art LLM, to directly evaluate the outputs of the student LLM utilizing the current library. This direct evaluation is especially coherent since the student has been trained on the teacher's strategies, making the teacher an ideal evaluator due to its understanding of the intricacies and subtleties required in the responses. We let the student interact with the environment to generate a validation set and let the teacher score it. This score is recorded and compared with the score from the previous iteration. The algorithm stops if there is no (significant) improvement after a pre-defined number of iterations (we set it to 2).

# 3.3. Deployment

The training algorithm generates a library comprising a set of scenarios and their associated strategy prompts. We will now discuss how this library is utilized during deployment.

As shown in Equation (7), the student's output is determined by the scenario, the base prompt, and the strategy prompt optimized for that specific scenario during training. However, during deployment, the prior conversations differ from the training scenarios, and there is no predefined  $p_{\text{strategy}}$  for the test input. To address this, we employ a method akin to retrieval-augmented generation (Lewis et al. 2020), where we identify the most similar scenario(s) from a library based on their embeddings. The corresponding strategy from these similar scenarios is then applied to the new, unseen input, i.e.,

student's output: 
$$\mathbf{a} = \mathcal{S}(\mathbf{s}|p_{\text{base}}, p_{\text{strategy}}(\tilde{\mathbf{s}})), \text{ where } \tilde{\mathbf{s}} = \underset{\mathbf{s}' \in \mathcal{L}}{\text{arg min }} d(\mathbf{s}, \mathbf{s}').$$
 (8)

By leveraging scenario embeddings, the library can efficiently match new, unseen inputs to the most similar scenarios, thereby providing a tailored strategy prompt.

Notably, the use of the library is flexible. When retrieving strategies, it is not necessary to limit retrieval to a single closest scenario; instead, one can retrieve k > 1 scenarios from the library, following the k-nearest neighbors approach. We provide a detailed analysis in Appendix A.2 and find that tuning k can further enhance performance.

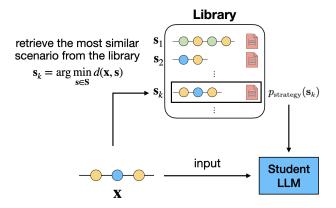


Figure 2 How the library is used during deployment

# 3.4. Method Summary and Discussion

Intuitively, the library consists of a set of representative and prototypical "reference" points within the data space. These reference points must *correctly* and *adequately* cover the data space to ensure that when encountering a new scenario, there is a high likelihood that an existing strategy can be effectively applied. The correct coverage necessitates the algorithm to prevent distribution shift, motivating the design of increasingly letting the student to generate scenarios in Section 3.2.1. The termination mechanism is in place to enforce adequate coverage, where the teacher LLM monitors the performance of the student and only stops when no improvement is observed, indicating the sufficiency of the coverage.

A key reason for the success of this method is the ability of the LLM to extrapolate, adapt, and apply a strategy designed for one scenario to a slightly different scenario. This capability allows the LLM to bridge gaps between similar but distinct situations, ensuring that the strategies remain effective even when faced with novel inputs. The LLM's inherent understanding of the nuances in language and context further strengthens its ability to generalize from the library's reference points, making it a powerful tool for dynamic and diverse real-world applications. This ability further allows the library to be used in a slightly different context where it is not trained on, as we will demonstrate in Section 4.3.

Interpretability Benefits. One of the defining features of our methodology is its interpretability: the teacher's knowledge is distilled into scenarios and prompts that are easily understood by humans. This interpretability offers several advantages over fine-tuningbased methods. First, domain experts can verify the library before deployment, which is especially important when the teacher is an external LLM over which the company has no control, including its model and training data. In fine-tuning methods, a student model can be easily contaminated under adversarial attacks, as it is unclear what knowledge has been distilled into the student. In contrast, our method keeps the student LLM's parameters unchanged, storing external knowledge in an interpretable library where each piece of "knowledge" can be scrutinized and verified. This design makes our method more resilient to adversarial attacks. Second, a library-based knowledge distillation is easily editable. Suboptimal or outdated strategies can be quickly updated by human experts. Meanwhile, when new scenarios arise during deployment and the existing library lacks relevant strategies, new entries can be seamlessly added with strategies devised by either the teacher or domain experts. This flexibility ensures that the library remains current and effective across various real-world applications.

# 4. Experiments

We evaluate our method on four student models, LlaMa-2 7b, LlaMa-2 13b, LlaMa-2 70b and GPT-3.5. We assess LlaMa models because they are open source and free except for storage and computing cost. We assess GPT-3.5 because it can only be queried by API and the cost of use is only 5% of GPT-4 at \$0.50 per 1M input tokens. This selection of student models also covers a large range of model size, from 7 billion (LlaMa-2 7b) to 175 billion (GPT-3.5).

With our experiments, we investigate the following research questions related to our proposed knowledge distillation method.

- Q1. Is teaching strategy more effective than teaching responses? To answer this question, for each task, we compare our proposed strategy teaching approach against fine-tuning student LLMs using the same amount of data. Fine-tuning directly trains the student on the scenario response pairs without explicitly teaching strategies like our method does.
- **Q2**. Is a context-specific dynamic use of strategies provided by the advanced LLM better than global strategies that are invariant across contexts? For this, we compare our dynamic approach against global strategies.
- Q3. How effective is our solution to fixing the distribution shift challenge? To answer this question, we create a baseline that works the same as ours with the only difference being that the scenarios are all generated by the teacher interacting with the customer.
- Q4. Are learnt strategies transferable across multiple student LLMs and other task contexts? We assess whether a strategy learnt from teaching one student LLM works for a different LLM and whether strategies learnt for one customer service context improves performance in another context.

### 4.1. Main Results

Baseline Methods. To answer question Q1 above, we compare our approach with traditional knowledge distillation via fine-tuning, where we first use the teacher LLM to interact with the customer LLM to generate a set of conversations, and then extract subconversations from it, just like how we sampled scenarios. The difference is that the student does not participate in the data generation and no explicit strategies are provided. The scenarios are used as the input and then we use the teacher's response as the output, creating a training dataset for fine-tuning. To answer Q2, we compare with a set of guidelines manually constructed, through iterations between the authors and GPT-4. These guidelines are global instructions that are used for every input the student encounters.

Evaluation Approach. We evaluate the LLMs at the conversation level in terms of the achievement of the pre-defined goal - customer satisfaction. Motivated by recent research (Fu et al., 2023; Chiang et al., 2023; Geng et al., 2023; Sun et al., 2023), we perform two sets of evaluations.

We first use advanced LLMs as an evaluator to rate each conversation from 1 (very dissatisfied) to 5 (very satisfied), using few-shot prompt where examples of rating 5 conversations are provided by the teacher, GPT-4, interacting with the customer LLM. We then use human evaluation where each participant does a blind comparison of GPT-4 and a student LLM, based on their conversations with the same customer LLM.

LLM as Evaluators. We use advanced LLMs to evaluate the conversations generated by different LLM agents (teacher or student) interacting with the customer LLM. Here we use GPT-4, since it is the state-of-the-art LLM available today. We use few-shot learning when evaluating these conversations - we include examples in the prompt that are generated by the teacher (GPT-4) and the customer. Since different examples may lead to different ratings, for robust evaluation, we perform the evaluation twice, using examples generated by the teacher interacting with "demanding" customers and "non-demanding" customers, respectively. We report the average ratings of four student models trained by different methods in Table 1.

Our strategy imitation method performs consistently better than the baselines. First, compared to global guidelines, our method completely dominates this baseline, performing better or equally in all experiments. Compared with fine-tuning, our method is better or equal except on the smallest LLM, LlaMa-2 7b. This is because smaller LLMs do not follow instructions as well as larger LLMs. Thus the knowledge distilled from the teacher may not be executed by the student as well as a larger student. This observation is consistent

Table 1 Average Customer Satisfaction Evaluated by GPT	Table 1	1 Average Custo	omer Satisfactio	n Evaluated b	y GPT-
--------------------------------------------------------	---------	-----------------	------------------	---------------	--------

Methods	Demanding Customer Conversation Baseline				Non-Demanding Customer Conversation Baseline			
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
BASE LLM STRATEGY IMITATION FINE-TUNING GLOBAL GUIDELINES	3.91 <b>4.88</b> 4.63 4.44	4.81 <b>5.0</b> 4.94 4.5	4.88 <b>5.0</b> <b>5.0</b> 4.75	4.91 <b>5.0</b> 4.68 <b>5.0</b>	2.84 3.44 <b>3.75</b> 3.31	3.59 <b>3.75</b> 3.47 3.44	4.03 <b>4.28</b> 4.06 3.88	4.03 <b>4.75</b> 3.81 3.75

with the findings from the literature that prompt based approaches are more effective for the LLM with more parameters (Lester et al. 2021).

Note that when prompts use conversations with demanding customers as examples of a rating of 5, they tend to yield higher scores across all evaluation tables. This occurs because when we label these challenging interactions as a 5 in the prompt, we communicate to the LLM that even when a customer is persistent and difficult to console, the conversation still merits a high score. This effectively sets a low bar for evaluation, leading to inflated ratings. Conversely, if examples feature non-demanding customers where positive interactions result in a rating of 5, the LLM is more likely to focus on the customer's reactions when evaluating conversations. This establishes a higher standard for ratings.

Different advanced LLMs can be used for evaluation. To assess robustness, we also use LlaMa-3 70B as an evaluator. See Appendix A.1 for details. In addition, we also evaluated retrieving more than 1 most similar scenarios to use their strategies during test. The results are shown in Appendix A.2.

Human Evaluation. We conduct a blind comparison of the teacher and student via human evaluation. We design a survey and show the two conversations, from the student and teacher, each interacting with the same customer. We then ask the participant to choose the better agent based on the conversations.

We report the percentage of times each student model is selected. If a student is comparable to the teacher, the selection rate should be around 50%. Thus, the larger the number,

the better the model. Results show that the base models for each student is consistently worse than the teacher, especially for LlaMa-2 7b, which is selected only 3% of the time. After the knowledge distillation, we see a consistent improvement in the performance of student LLMs, with weaker LLMs benefiting more from learning from the teacher LLM. While the student still performs slightly worse than the teacher, the precentages significantly increase.

Table 2 The percentage of times the student agent is selected over GPT-4 by human subjects

Methods	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
Base LLM	0.03	0.25	0.38	0.22
STRATEGY IMITATION	0.34	0.41	0.41	0.41

Discussion. Overall, our method performs better than the global guidelines because it provides strategies customized for each scenario and for each student. These strategies are developed based on specific scenarios, such that dynamic instructions are provided to the student, in contrast with the fixed strategies from a global guideline. Consequently, vastly different, even contradictory strategies may exist. We shown an example in Figure 3. In a scenario where a customer initiates a conversation by asking for help, the student is instructed to avoid over-explaining policies. In contrast, in a different scenario where the customer is engaged and seeks information, the student is instructed to provide clear and detailed information. Conversely, the global guidelines remain fixed, lacking such flexibility and adaptability.

In addition, strategies are customized for different students. Strategies are constructed via proposing changes to the students' current output. In other words, the strategies are "delta" to the students' output, which is targeted to fix the problem of the student. Different student LLMs behave differently, and thus will receive different strategies from the

### Scenario 1: beginning of a conversation

CUSTOMER: Hello, I hope you're doing well. I recently booked a flight with your airline, but I need to cancel it. I understand that it's a non-refundable and non-changeable ticket, but I was hoping you could make an exception in this case and provide me with a full refund.

#### CUSTOMER: Hello, I recently booked a flight with your airline, and I would like to inquire about canceling it... AGENT: Hello, thank you for reaching out to us. ...

Scenario 2: customer engaged and asking for information

CLISTOMER:

AGENT: ... I hope one of these options might be suitable for you. Please let me know how you'd like to proceed

CUSTOMER: Thank you for providing me with this information... Could you please provide me with more details on how these travel credits work and how I can use them for future travel with your airline?

# Strategy Strategy · Avoid Over-explaining Policies: While it's important to explain Provide Clear and Detailed Information: Provides a clear and policies, avoid excessive detail that may overwhelm or confuse the customer. detailed explanation regarding the policy

Figure 3 Different strategies are learned for different scenarios.

teacher. For example, LlaMa2-7b tends to generate longer responses than GPT-3, the average response length is about 50% more than that of GPT-3.5. We calculate the percentage of strategies that contain keywords "concise", "brevity" or "excessive" in it, which means the teacher instructs the student to provide shorter responses. We find that 15.4\% of the strategies for GPT-3.5 contain such keywords while 22.1% of strategies for LlaMa 2-7b contain such keywords, a 50% increase.

Therefore, our method design ensures the strategies are customized to each student as well as each scenario the student might encouter during deployment.

# Evaluating the Solution to Distribution Shift

To answer question Q3, we compare our method with a baseline solution where everything else remains the same as in our proposed method, except that we do not address the distribution shift. In the baseline, which we call "Strategy Imitation (ablated)", all scenarios are generated by only the teacher interacting with the customer, rather than allowing the student to participate in the generation process. To ensure a fair comparison, we generate the same number of scenarios as the original algorithm. In this case, a distribution shift exists because the training data is generated by the teacher, while during deployment, the scenarios are generated by the student.

To assess the distribution shift quantitatively, we report the distance between a new scenario the student encounters during deployment and the closest scenario retrieved from the library. The strategy from the retrieved scenario will be used to instruct the student on handling the new scenario. Therefore, the distance between the new and retrieved scenario determines the applicability of the strategy to the new scenario. A small distance indicates that the retrieved scenario is very similar to the new scenario, making the strategy highly relevant and applicable. Conversely, a larger distance suggests that the scenario is less relevant, leading to a lower quality of the corresponding strategy.

Table 3 Average scenario distances

METHODS	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
STRATEGY IMITATION	0.179	0.172	0.181	0.169
STRATEGY IMITATION (ABLATED)	0.192	0.192	0.191	0.174
% Increase	7.3%	11.6%	5.5%	3.0%

Table 3 presents the average distances between new scenarios and retrieved scenarios for all conversations in the test set, compared to those of the original algorithm. The results show that the average distance increased by up to 12%, indicating that the retrieved scenarios are relatively less similar to the test cases compared to the original algorithm. This is because the distribution of the test data deviates more from the training data.

Table 4 evaluates the performance of the ablated algorithm using GPT-4. Without handling the distribution shift, the ablated algorithm performs worse than the original algorithm. Occasionally, the ablated algorithm also performs worse than the base LLM, as the strategies become less relevant when the distances increase, which means the student follows bad advice on how to respond to customers.

We provide more analyses to show the effect of distribution shift in Appendix B.

Table 4 Average Customer Satisfaction Evaluated by GPT-4

Methods	Demanding Prompt				Non-Demanding Prompt			
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
BASE LLM STRATEGY IMITATION STRATEGY IMITATION (ABLATED)	3.91 <b>4.88</b> 4.5	4.81 <b>5.0</b> 4.75	4.88 <b>5.0</b> 4.63	4.91 <b>5.0</b> 4.81	2.84 <b>3.44</b> 3.25	3.59 <b>3.75</b> 3.56	4.03 <b>4.28</b> 4.06	4.03 <b>4.75</b> 4.38

# 4.3. Transferrability

We have shown that the library is customized for each student and takes into account their specific characteristics and current ability in responding to the customers. Therefore, the library should contain different strategies for different students. On the other hand, we also hypothesize that there may exist some knowledge that is common to all students, even to other different but related scenarios. To study this question, we test the *transferrability* of the library.

An important advantage of our method is that the library is highly portable and reusable; other student LLMs can directly use it without training. Therefore, we let all LLMs use the library built for LlaMa-2 7b, the LLM with the lowest cost, instead of training their own library.

Table 5 Model Transferability Assessment (Evaluated by GPT-4)

METHODS	Demanding Prompt				Non-Demanding Prompt			
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
Base LLM	3.91	4.81	4.88	4.91	2.84	3.59	4.03	4.03
STRATEGY IMITATION	4.88	5.0	5.0	5.0	3.44	3.75	4.28	4.75
LlaMa-2 7b Library	4.88	5.0	4.88	5.0	3.44	3.91	3.94	4.44

Results show that the transferred library improves performance from the base LLM, even though it is not directly trained for other students. This indicates the student LLMs may lack some common communication skills and thus benefit from strategies constructed for other LLMs.

The transferrability of the library is not only across LLMs, but also across contexts. Though the library is built for a specific context of "canceling a restricted ticket," we hypothesize that some aspects of the knowledge on how to interact with customers should also apply to other contexts within the general domain of customer service.

Therefore, we evaluate the performance of our approach in a new context outside of the training. Specifically, we consider a context, where customers call the airline to resolve a "lost luggage issue," where the customer seeks \$10,000 compensation for their loss. For this experiment, we again compare performance using our approach with fine-tuning and global guidelines. For fine-tuning, we directly use the LLM fine-tuned on the ticket problem to solve the luggage problem without retraining. For the global guidelines, we use the same global guidelines constructed for the ticket problem. The results are shown in Table 6.

Table 6 Context Transferability Assessment Evaluated by GPT-4

Methods	Demanding Prompt				No-Demanding Prompt			
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
Base LLM	4.5	5.0	5.0	4.91	3.13	4.25	4.5	4.63
STRATEGY IMITATION	4.5	5.0	5.0	5.0	4.0	4.63	4.5	5.0
FINE-TUNING	5.0	5.0	4.88	5.0	4.5	4.63	4.63	4.75
Global Guidelines	4.75	5.0	4.88	5.0	3.88	4.25	4.63	5.0

In addition, we also test using strategies from more than one most similar scenario, specifically k = 5, and we find the results are even better than k = 1, as discussed in Appendix A.2.1.

From the analyses, we find that the library is transferable across contexts, providing superior performance. The transferability is better than fine-tuning when the student LLM has more parameters (e.g., LlaMa-2 70B and GPT-3.5). For smaller LLMs, fine-tuning performs better. We hypothesize that this is because larger LLMs are better at extrapolating and may inherently adapt the strategies to their own context, even if the strategies initially appear less relevant.

# 5. Conclusion

Though advancement in LLMs have broadened their applicability in various domains, including customer service and content generation, deploying these models often faces constraints related to security, their size, cost, and operational feasibility on resource-constrained devices. Researchers have developed methods to enhance smaller, more economical models that require fewer computational resources for real-time deployment to address these constraints.

This paper introduces "interpretable knowledge distillation," a novel approach for transferring strategic knowledge from a large, sophisticated "teacher" LLM to a smaller "student" LLM. Unlike traditional methods that rely on direct imitation from teacher responses for learning, our approach focuses on strategy teaching. The teacher model provides strategies for handling various scenarios, which are compiled into a library. This library serves as a reference for the student model to apply appropriate strategies in real-world situations, functioning similarly to retrieval-augmented generation (RAG). This method does not require modifying the student model's parameters, allowing it to benefit from the teacher's knowledge through external, dynamic referencing, thus ensuring adaptability and scalability.

Experimental results validate the method's effectiveness. In a customer service context, student models equipped with the strategy library consistently outperformed those trained with traditional fine-tuning methods while gaining interpretability. They achieved higher customer satisfaction rates and effectively adapted to new, unseen scenarios. Transferability tests showed that strategies learned in one context (e.g., ticket cancellation) were beneficial in another (e.g., lost luggage claim), highlighting the method's broad applicability. Moreover, the approach's interpretability facilitates human oversight and ongoing

refinement of the strategy library. The results suggest that smaller LLMs can achieve near parity with their larger counterparts in specific tasks, reducing the need for higher-cost, higher-complexity models. We hope the various advantages of our interpretable knowledge distillation approach facilitates greater adoption of LLMs across a range of challenging marketing tasks.

# Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

### References

- Agarwal R, Vieillard N, Zhou Y, Stanczyk P, Garea SR, Geist M, Bachem O (2024) On-policy distillation of language models: Learning from self-generated mistakes. The Twelfth International Conference on Learning Representations.
- Agarwal R, Vieillard N, Zhou Y, Stanczyk P, Ramos S, Geist M, Bachem O (2023) Generalized knowledge distillation for auto-regressive language models. arXiv preprint arXiv:2306.13649.
- Bordes A, Boureau YL, Weston J (2016) Learning end-to-end goal-oriented dialog.  $arXiv\ preprint$  arXiv:1605.07683.
- Brand J, Israeli A, Ngwe D (2023) Using gpt for market research. Available at SSRN 4395751.
- Cheng Y, Liu W, Wang J, Leong CT, Ouyang Y, Li W, Wu X, Zheng Y (2024) Cooper: Coordinating specialized agents towards a complex dialogue goal. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17853–17861.
- Deng Y, Liao L, Liang C, Hongru W, Lei W, Chua TS (2023) Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. The 2023 Conference on Empirical Methods in Natural Language Processing.
- Gui G, Toubia O (2023) The challenge of using llms to simulate human behavior: A causal inference perspective.  $arXiv\ preprint\ arXiv:2312.15524$ .
- Ham D, Lee JG, Jang Y, Kim KE (2020a) End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. Jurafsky D, Chai J, Schluter N, Tetreault J, eds., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 583-592 (Online: Association for Computational Linguistics), URL http://dx.doi.org/10.18653/v1/2020.acl-main.54.
- Ham D, Lee JG, Jang Y, Kim KE (2020b) End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. Proceedings of the 58th annual meeting of the association for computational linguistics, 583–592.
- Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network.  $arXiv\ preprint$  arXiv:1503.02531.

- Horton JJ (2023) Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.
- Hsieh CY, Li CL, Yeh Ck, Nakhost H, Fujii Y, Ratner A, Krishna R, Lee CY, Pfister T (2023) Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes.

  Findings of the Association for Computational Linguistics: ACL 2023, 8003–8017.
- Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
- Joshi R, Balachandran V, Vashishth S, Black A, Tsvetkov Y (2021) Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. *International Conference on Learning Representations (ICLR)*.
- Kim Y, Rush AM (2016) Sequence-level knowledge distillation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1317–1327.
- Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. *Proceedings*of the 2021 Conference on Empirical Methods in Natural Language Processing, 3045–3059.
- Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih Wt, Rocktäschel T, et al. (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33:9459–9474.
- Li P, Castelo N, Katona Z, Sarvary M (2024) Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science* .
- Li Y, Zhang Y, Sun L (2023) Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents.  $arXiv\ preprint\ arXiv:2310.06500$ .
- Magister LC, Mallinson J, Adamek J, Malmi E, Severyn A (2022) Teaching small language models to reason.  $arXiv\ preprint\ arXiv:2212.08410\ .$
- Merrill DW, Reid RH (1981) Personal styles & effective performance (CRC Press).
- Pomerleau DA (1991) Efficient training of artificial neural networks for autonomous navigation. *Neural computation* 3(1):88–97.

- Qiu L, Singh PV, Srinivasan K (2023) How much should we trust llm results for marketing research? Available at  $SSRN\ 4526072$ .
- Ross S, Bagnell D (2010) Efficient reductions for imitation learning. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 661–668 (JMLR Workshop and Conference Proceedings).
- Samad AM, Mishra K, Firdaus M, Ekbal A (2022) Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. Findings of the Association for Computational Linguistics: NAACL 2022, 844–856.
- Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.  $arXiv\ preprint\ arXiv:1910.01108$ .
- Snell C, Yang S, Fu J, Su Y, Levine S (2022) Context-aware language modeling for goal-oriented dialogue systems. Findings of the Association for Computational Linguistics: NAACL 2022, 2351–2366.
- Sun S, Cheng Y, Gan Z, Liu J (2019) Patient knowledge distillation for bert model compression. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 4323–4332.
- Tang R, Lu Y, Liu L, Mou L, Vechtomova O, Lin J (2019) Distilling task-specific knowledge from bert into simple neural networks. arXiv preprint arXiv:1903.12136.
- Wang X, Shi W, Kim R, Oh Y, Yang S, Zhang J, Yu Z (2019) Persuasion for good: Towards a personalized persuasive dialogue system for social good. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5635–5649.
- Wei W, Le Q, Dai A, Li J (2018) Airdialogue: An environment for goal-oriented dialogue research. *Proceedings*of the 2018 Conference on Empirical Methods in Natural Language Processing, 3844–3854.
- Zhang H, Zeng Z, Lu K, Wu K, Zhang S (2022) Efficient dialog policy learning by reasoning with contextual knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11667–11675.
- Zhang Q, Naradowsky J, Miyao Y (2023) Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. Findings of the Association for Computational Linguistics:

  ACL 2023, 6665–6694.

# **Appendix**

#### A. Robustness Checks

### A.1. Alternative LLM for Evaluation

In the main paper, we used GPT-4 to evaluate the model. To assess robustness with respect to the LLM that evaluates the agent, we use LlaMa 3-70b in place of GPT-4 and follow the same procedure in Section 4.1. The results are shown in Table 7. The evaluations are consistent with those from GPT-4 in Table 1: the average correlation between GPT-4 and LlaMa 3-70b evaluations is 0.81. With both LLM evaluators, our proposed method performs consistently better than the baselines, only occasionally worse than fine-tuning, while gaining interpretability.

Table 7 Average Customer Satisfaction Evaluated by Llama 3-70b

Methods	Demanding Prompt				Non-Demanding Prompt			
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
Base LLM	3.97	4.69	4.81	4.88	3.88	4.31	4.38	4.63
STRATEGY IMITATION	4.69	4.94	4.94	5.0	4.13	4.5	4.31	4.72
Fine-Tuning	4.63	4.72	4.88	4.94	4.44	4.5	4.5	4.41
Global Guidelines	4.38	4.13	4.69	5.0	3.91	3.84	4.13	4.72

### A.2. Using kNN to Retrieve Guidelines to Improve Performance

During deployment, our method uses the library by retrieving the most similar scenario and uses the corresponding strategy to instruct the student LLM to respond to the customer. This mechanism can be generalized following the idea of k-nearest neighbor. Instead of retrieving one most similar scenario, our method can be generalized to use k most similar scenarios, allowing the algorithm to be further tuned to improve the performance. Here we perform additional experiments of using k = 5 and use GPT-4 to evaluate the conversations, to compare with the one generated by the original algorithm where k = 1. The results are shown in Table 8. Note that when aggregating strategies from multiple scenarios (e.g., when k = 5), the list could can exceed the input token limit. In such cases, we use GPT4 to summarize the guidelines into a shorter list before being used by the student.

Results show that k does have an impact on the performance of LLM as larger k indicates a longer and richer list of guidelines. However, it is interesting that more instructions do not always improve performance; the improvement depends on the context and the LLM. Therefore, in practice, we recommend that k is treated as a hyper-parameter and tuned via a validation set before deployment. Tuning of k can be independent of the training step, and can be customized for different use of the library.

Table 8 Evaluation of Conversations Generated by Retrieving Strategies of 1 or 5 Most Similar Scenarios

Methods		Demandin	g Prompt			Non-Demand	ling Prompt	
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 $70b$	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
K=1	4.88	5.0	5.0	5.0	3.44	3.75	4.28	4.75
K = 5	4.75	4.75	5.0	5.0	3.66	4.03	4.06	4.66

A.2.1. How use of kNN to Retrieve Guideliness Improve Context Transferability We also tested using a larger k, specifically k = 5, when retrieving guidelines across different contexts. As shown in Table 9, the evaluation by GPT-4 indicates that a larger k tends to lead to better performance for cross-context use of the library. This is because, in a new context, every scenario in the library is relatively distant from the input. Therefore, there is not much difference in relevance between the most similar (closest) scenario and the k-th closest scenario. Including more scenarios does not negatively impact relevance, but it increases the diversity of advice, which can ultimately benefit the model.

Table 9 Context Transferability Assessment Evaluated by GPT-4 for different K

Methods		DEMANDING	G PROMPT		No-Demanding Prompt			
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5
Base LLM K=1 K=5	4.5 4.5 <b>4.88</b>	5.0 5.0 5.0	5.0 5.0 5.0	4.91 <b>5.0</b> <b>5.0</b>	3.13 <b>4.0</b> <b>4.0</b>	4.25 <b>4.63</b> 4.6	4.5 4.5 <b>4.75</b>	4.63 <b>5.0</b> <b>5.0</b>

# B. Distribution Shift Drives Drop in Performance: Evidence

In Section 4.2, to demonstrate the value of preventing distribution shift in our algorithm, we ran an ablated version of our algorithm. In the ablated algorithm, we do not specifically tackle distribution shift - all scenarios are generated only by the teacher interacting with the customer LLM. The evaluation in Table 4 shows that the ablated algorithm downgrades the performance significantly.

To show that the decrease of performance is indeed due to distribution shift, that the later scenario in the conversation, the more deviation there will be in the algorithm, we conduct another evaluation. We use GPT-4 to evaluate only the first half of the conversation generated by the original algorithm and the ablated algorithm, interacting with the same customer LLM with the same prompt. We then compute the drop of average rating in percentage in Table 10, which is the difference in the rating devided by the rating of the original algorithm, denoted as  $\Delta_{\text{half}}$  in the Table. We also compute the drop in rating from the full conversations evaluated in Table 4, denoted as  $\Delta_{\text{full}}$ . If there is distribution shift, then  $\Delta_{\text{half}}$  should be a lot

smaller than  $\Delta_{\text{full}}$  since there should be more deviation at the end of the conversations than at the beginning of the conversations.

Results show that  $\Delta_{\text{half}}$  is always less than half of  $\Delta_{\text{full}}$ , indicating the deviation in performance does not happen evenly across a conversation but more towards the end.

Table 10 Evaluation of Part of Conversations

Methods		Demandin	g Prompt		Non-Demanding Prompt				
	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	LlaMa-2 7b	LlaMa-2 13b	LlaMa-2 70b	GPT-3.5	
STRATEGY IMITATION	4.88	5.0	5.0	5.0	4.85	4.91	4.97	4.88	
STRATEGY IMITATION (ABLATED)	4.91	4.88	4.88	4.97	4.72	4.88	4.88	4.72	
$\Delta_{ m half} \ \Delta_{ m full}$	-0.6% -7.8%	-2.4% -5.0%	-2.4% -7.4%	-0.6% -3.8%	-2.7% -5.5%	-0.6% -5.1%	-1.8% -5.1%	-3.3% -7.8%	

# C. How Teacher LLM Iteratively Updates Strategies for Student LLM: Example

As shown in Figure 1, the strategies for each scenario are learned iteratively by comparing the student's updated response with the teacher's response and proposing changes to the strategies. As the strategies are updated, the student iteratively refines its response to become more and more similar to the teacher's response. Below we illustrate the iterative process using an example.

In the example below, the scenario consists of two turns. This exchange represents the core of the conversation, where the initial pleasantries and information gathering are complete. At this stage, the agent transitions to the critical task of providing the customer with the necessary assistance and solutions.

### Scenario

Agent: Hello, how can I help you?

Customer: Hello, I recently booked a flight with your airline, and I need to cancel it. I tried to do it online, but I couldn't find an option to cancel. Can you assist me with canceling my flight?

**Agent**: Of course, I'd be happy to assist you with that. Could you please provide me with your booking reference number so I can pull up your reservation details?

Customer: Sure, my booking reference number is ABC123.

The student's initial response (from the base LLM, without following any strategies from the teacher) is very different from the teacher's response. The student does not provide any explanation of the policy or offer alternative solutions but directly expresses rejection. It overall sounds very cold and blunt.

### Teacher Response

Teacher: Thank you for providing the booking reference number. Please allow me a moment to retrieve your booking details. [After a moment] I see that you have booked a restricted ticket. As per our policy, restricted tickets are non-refundable and non-changeable. This is because these tickets are offered at a lower cost compared to our flexible tickets. I understand this might not be the news you were hoping for, and I apologize for any inconvenience this may cause. However, we do have some alternative solutions that might be helpful. We can offer you travel credits for the value of your ticket, or the possibility of a one-time rescheduling, subject to applicable fees. Would you like to hear more about these options?

### Student Response

**Student (base LLM)**: Thank you for providing that information. I have pulled up your reservation, and I see that you have a restricted ticket, which is non-refundable and non-changeable. I understand that you wish to cancel your flight, but I must inform you that as per the purchase agreement, restricted tickets are not eligible for free cancellations or refunds.

Based on these two responses, the teacher compares and then proposes changes to the student to improve, using the teacher's response as the desired output. The teacher's strategies at the first iteration point out the issues and offer suggestions for improvement. Incorporating the suggestions, the student's response after iteration 1 improved significantly. The student offered an explanation of the policy and offered alternative solutions to the customer.

### Strategies and Updated Student Responses At Iteration 1

### **Strategies:**

- 1. Acknowledge Customer's Request: Acknowledge the customer's request promptly and express willingness to assist.
- 2. Provide Explanation: Offer clear and empathetic explanations of the situation, such as ticket restrictions.
- 3. Offer Alternatives: Demonstrate a proactive approach by providing alternative solutions or options to assist the customer.
- 4. Avoid Blunt Statements: Avoid using blunt statements without context or empathy, as it can be dismissive to the customer.
- 5. Avoid Lack of Explanation: Provide clear explanations to avoid leaving the customer confused or frustrated.
- 6. Avoid Failure to Offer Solutions: Offer alternative solutions or options to prevent a negative customer experience.

Student: Thank you for providing your booking reference number, ABC123. After reviewing your reservation, I see that you have purchased a restricted ticket. I understand that you wish to cancel your flight; however, I must inform you that restricted tickets are non-refundable and non-changeable as per the purchase agreement. This policy allows us to offer these tickets at a lower cost. While I cannot provide a full refund for your ticket, I can offer you some alternatives. You may be eligible for travel credits or a one-time rescheduling of your flight, subject to applicable fees. Would you like to explore these options further?

This process continues as the strategies are updated over iterations and the student further improves its response, until it sounds "close enough" to the teacher's response. The final strategies and the student response are shown below. These strategies are then paired with the scenario and stored in the library, and included in the prompt during deployment.

### Strategies and Student Responses at the Last Iteration

### **Strategies:**

- 1. Acknowledge Customer's Request: Promptly acknowledge the customer's request and express willingness to assist.
- 2. Provide Explanation: Offer clear and empathetic explanations of policies or restrictions that may affect the customer's request.
- 3. Offer Alternatives: Demonstrate a proactive approach by providing alternative solutions or options to assist the customer.
- 4. Express Empathy and Understanding: Show empathy and understanding towards the customer's situation.
- 5. Maintain Professional and Courteous Tone: Use polite language and maintain a professional tone throughout the conversation.
- 6. Avoid Blunt or Overly Formal Language: Refrain from using blunt or overly formal language that may sound robotic or impersonal.
- 7. Avoid Lack of Empathy: Ensure empathy and acknowledgment of customer disappointment in responses.
- 8. Avoid Transactional Communication: Be proactive in offering options and solutions rather than being too transactional in responses.

Student: AGENT: Thank you for providing your booking reference number, ABC123. After reviewing your reservation, I see that you have purchased a restricted ticket. I regret to inform you that restricted tickets are non-refundable and non-changeable as per the purchase agreement. This policy allows us to offer these tickets at a lower cost. I understand that you need to cancel your flight, and I empathize with your situation. While I cannot provide a full refund for the restricted ticket, I can offer you some alternative solutions. One option is to consider converting the value of your ticket into travel credits that you can use for future bookings with our airline. Additionally, if you need to reschedule your flight, we can explore the possibility of a one-time rescheduling, subject to applicable fees. I hope these alternatives can be helpful to you. Please let me know how you would like to proceed, and I will assist you accordingly.