AI-Salesman: Towards Reliable Large Language Model Driven Telemarketing

Qingyu Zhang[1, 2*], Chunlei Xin[1, 2*], Xuanang Chen[1], Yaojie Lu[1†], Hongyu Lin[1†], Xianpei Han[1, 2], Le Sun[1, 2], Qing Ye[3], Qianlong Xie[3], Xingxing Wang[3]

[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Independent Researcher
zhangqingyu2024@iscas.ac.cn, chunlei2021@iscas.ac.cn

Abstract

Goal-driven persuasive dialogue, exemplified by applications like telemarketing, requires sophisticated multi-turn planning and strict factual faithfulness, which remains a significant challenge for even state-of-the-art Large Language Models (LLMs). A lack of task-specific data often limits previous works, and direct LLM application suffers from strategic brittleness and factual hallucination. In this paper, we first construct and release TeleSalesCorpus, the first real-world-grounded dialogue dataset for this domain. We then propose AI-Salesman, a novel framework featuring a dual-stage architecture. For the training stage, we design a Bayesian-supervised reinforcement learning algorithm that learns robust sales strategies from noisy dialogues. For the inference stage, we introduce the Dynamic Outline-Guided Agent (DOGA), which leverages a pre-built script library to provide dynamic, turn-by-turn strategic guidance. Moreover, we design a comprehensive evaluation framework that combines fine-grained metrics for key sales skills with the LLM-as-a-Judge paradigm. Experimental results demonstrate that our proposed AI-Salesman significantly outperforms baseline models in both automatic metrics and comprehensive human evaluations, showcasing its effectiveness in complex persuasive scenarios.

Datasets —
https://huggingface.co/datasets/ICIP/TeleSalesCorpus

1 Introduction

While conversational AI has made significant strides in both structured task-oriented dialogue (Ham et al. 2020; Hosseini-Asl et al. 2020; Xu et al. 2024) and unconstrained open-domain chit-chat (Gao, Galley, and Li 2018; Roller et al. 2021; Friedman, Panigrahi, and Chen 2025), a critical and challenging frontier remains underexplored: goal-driven persuasive dialogue for intelligent marketing, unlike conventional dialogue tasks, intelligent marketing, exemplified by telemarketing, requires conversational AI to actively strategize, persuade, and guide users toward specific outcomes. This presents a unique confluence of high-stakes challenges that current large language models (LLMs) struggle to address effectively.

The core challenges of intelligent telemarketing are threefold. First is the challenge of satisfaction. The AI must not only generate human-like responses but also navigate a wide variety of marketing scenarios, each with its own complex strategies and logical flows. General-purpose large language models, despite their fluency, struggle to capture and reliably execute these diverse, long-horizon conversational plans (Valmeekam et al. 2023; Pan et al. 2025; Chen et al. 2025a; Lin et al. 2025), failing to satisfy the strategic requirements of the task. Second is the challenge of faithfulness. In high-stakes sales interactions, the AI must adhere strictly to the constraints of the product or service. However, the propensity of LLMs for factual hallucination (Maynez et al. 2020; Rawte et al. 2023; Atanasova et al. 2023; Chen et al. 2025b,c) poses an unacceptable risk, potentially resulting in misleading claims or inaccurate commitments. Third is the challenge of customization. Each customer possesses a unique background, with distinct concerns and points of interest. Effective persuasion requires tailoring arguments and information delivery to individual needs. Yet, LLMs frequently produce generic responses and lack the strategic reasoning necessary to address specific objections effectively (Fu et al. 2023).

To address these multifaceted challenges, this paper introduces AI-Salesman, an end-to-end framework that tackles these issues through innovations at both the training and inference stages, as illustrated in Figure 1. Specifically, AI-Salesman integrates two core mechanisms to achieve this. First, to satisfy the critical demands of satisfaction and faithfulness, we introduce a novel reward function grounded in Bayesian principles into our Group Relative Policy Optimization (GRPO) training process (Shao et al. 2024). Moving beyond conventional outcome-based rewards, our approach directly supervises the model's intermediate reasoning. Inspired by Bayesian principles, we decompose the reward signal for a thought process into two intuitive criteria: a prior that captures the intrinsic coherence of the reasoning itself, and a likelihood that measures its strategic utility in justifying the expert's final response. By optimizing for both coherent reasoning and effective outcomes, the model learns to generate responses that are both factually grounded and persuasive, thereby enhancing user satisfaction and faithful-

*These authors contributed equally.
†Corresponding author.

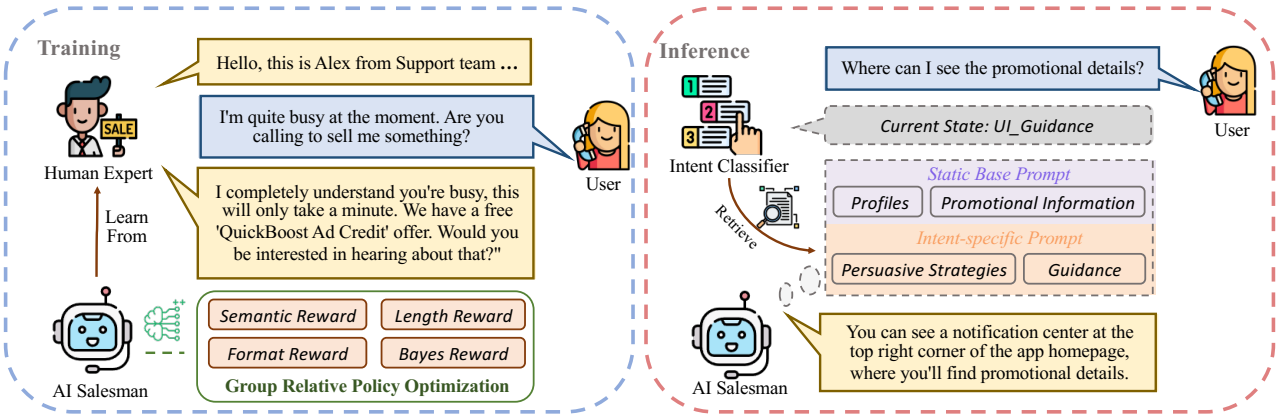


Figure 1: Overview of Training and Inference for the AI Salesman.

ness. Second, to enable customization, we propose the Dynamic Outline-Guided Agent (DOGA), a framework that operates during the inference stage. To overcome the generic responses common with static prompting, DOGA dynamically constructs a tailored strategy outline for each turn. By analyzing the user's profile, real-time intent, and dialogue history, it retrieves the most relevant persuasive strategies from a pre-verified library. This curated outline then guides the LLM, ensuring its responses are strategically targeted to each customer's unique concerns and objections.

Unfortunately, a significant barrier to progress in this domain is the absence of specialized training data and effective evaluation methods for telemarketing (He et al. 2018; Wang et al. 2019). To address this gap, we first introduce TeleSalesCorpus, a large-scale corpus of high-fidelity dialogues generated through a state-aware simulation grounded in real-world expert interactions. This corpus captures the complex patterns, customer objections, and conversational nuances characteristic of authentic sales conversations. Second, moving beyond simplistic success metrics, we propose a comprehensive evaluation framework specifically designed for telemarketing to enable fine-grained analysis. To systematically assess a model's ability to achieve strategic satisfaction, maintain factual faithfulness, and deliver persuasive customization, we define six sales capabilities, ranging from Business Analysis to Objection Handling, each assessed using a detailed rubric composed of seven qualitative metrics. By integrating this structured evaluation schema with the LLM-as-a-Judge paradigm (Zheng et al. 2023; Chan et al. 2024), our framework supports rigorous and comprehensive assessment of model performance across diverse scenarios. This evaluation approach provides a scalable offline alternative to resource-intensive online A/B tests.

Overall, our contributions can be summarized as follows:

- We propose AI-Salesman, a novel end-to-end framework that integrates reasoning-aware reinforcement learning with dynamic outline-guided inference. To the best of our knowledge, this is the first LLM-based framework specifically designed for real-world telemarketing that systematically addresses the challenges of satisfaction, faithfulness, and customization.
- We construct and release TeleSalesCorpus, the first large-scale, high-fidelity dialogue dataset grounded in real-world sales conversations, specifically designed for training and evaluating telemarketing models.
- We propose a comprehensive offline evaluation framework across six core sales capabilities, enabling efficient and rigorous assessment of models' practical sales proficiency in diverse scenarios.

2 Telemarketing Scenarios

To systematically analyze model performance in telemarketing, this section addresses two key aspects. First, we formally define the dialogue generation task to articulate its underlying structure. Second, we introduce a comprehensive framework designed to evaluate the model performance across critical sales capabilities and qualitative metrics.

2.1 Task Definition

We model telemarketing dialogue as a conditionally constrained sequence generation task. At each turn t, the model generates a response based on the system prompt $\mathcal{P}$ and the dialogue history $\mathcal{H}_t = \mathcal{H}_{t-1} \oplus U_t$, where U_t is the user's utterance at turn t. The prompt $\mathcal{P}$ defines the task's global context, including a set of goals $G = \{g_1, \ldots, g_n\}$ and constraints $C = \{c_1, \ldots, c_m\}$.

The model's objective is to generate a response sequence A_t that maximizes its conditional probability given the inputs $(\mathcal{P}, \mathcal{H}_t)$. Formally, we seek the optimal response A_t^*:

$$A_t^* = \arg\max_{A_t \in \mathcal{V}^*} P(A_t | \mathcal{P}, \mathcal{H}_t) \quad (1)$$

where $\mathcal{V}$ is the model's vocabulary and $\mathcal{V}^*$ denotes its Kleene closure, representing the set of all possible sequences the model can generate.

This generation is subject to two primary conditions. First, the response A_t must adhere to all predefined rules, such that for every constraint $c \in C$, the condition $c(A_t) = 1$ is satisfied. Second, the response must be goal-oriented, designed to maximize the expectation of achieving the final task goals defined in G.

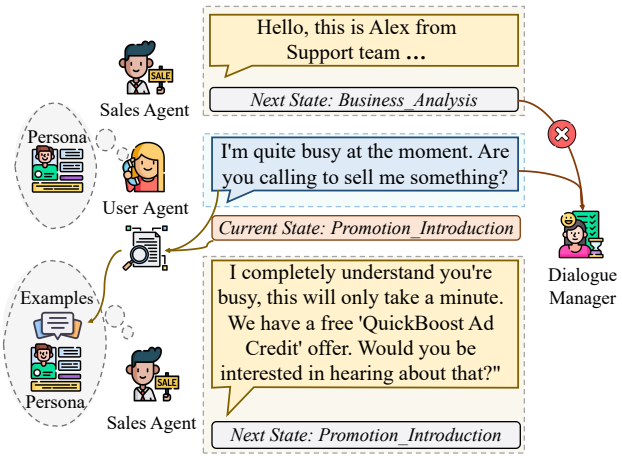


Figure 2: Data Construction Framework Overview.

2.2 Evaluation Framework

Our evaluation framework is built upon two core components: six fundamental sales capabilities required for the task and a rubric of seven evaluation metrics for granular, turn-by-turn assessment.

The six capabilities cover the entire lifecycle of a sales call: Role-playing, Business Analysis, Activity Introduction, Idle-chat Rejection, Objection Handling, and Operational Guidance. To provide a fine-grained assessment across these capabilities, we evaluate each response using seven qualitative metrics: Guideline Adherence(**Gui.**), Factual Correctness(**Fac.**), Logical Coherence(**Log.**), User Need Fulfillment(**Use.**), Response Richness(**Res.**), Safety(**Saf.**), and Completeness(**Com.**).

To operationalize this framework at scale, we employ GPT-4 as a judge. For each dialogue, the LLM-judge is given the conversation history, ground-truth data, and our metric definitions. Then it synthesizes these inputs to generate a holistic quality score on a 1-10 scale. This approach enables nuanced, context-aware evaluation that approximates human judgment for robustly benchmarking different models.

3 End-to-End Intelligent Sales System

3.1 Data Construction

The availability of suitable training data fundamentally constrains the development of a robust, goal-oriented persuasive dialogue system. Existing datasets (Wang et al. 2019; He et al. 2018) do not adequately address the unique challenges of telemarketing, such as complex business rules and specific promotional objectives. To bridge this gap, we constructed TeleSalesCorpus, a dataset using a semi-synthetic framework that leverages real-world expertise to generate high-fidelity, goal-oriented dialogues.

Our data creation process employs a state-aware, three-agent simulation, as illustrated in the Figure 2 provided. The framework features a User Agent with a distinct persona, a Sales Agent responsible for persuasion, and a central Dialogue Manager that orchestrates the interaction. At each turn, when the User Agent responds, the Dialogue Manager intervenes. It first adjudicates the true conversational state, overriding incorrect state predictions from the sales agent. Then, it queries a pre-compiled library of real-world interaction examples, retrieving a strategically relevant example based on the current state. This example is used to dynamically guide the Sales Agent in crafting a response that is both contextually appropriate and strategically sound.

This process is grounded in assets distilled from real dialogues and diverse, LLM-authored business scenarios. Following a rigorous, multi-faceted quality assurance protocol, our pipeline produced a final dataset of 2,000 high-fidelity conversations.

3.2 Stage-1: GRPO Training

To address the core challenges of satisfaction and faithfulness in intelligent telemarketing, we propose a policy optimization framework that synergizes the Group Relative Policy Optimization (GRPO) algorithm(Shao et al. 2024) with a novel Bayesian-Supervised Reasoning reward. GRPO facilitates online exploration of sales strategies, enabling the model to learn robust policies from noisy data. This exploration is guided by our Bayesian reward, which uniquely assesses the model's intermediate reasoning process. It assigns a higher value to reasoning that provides a logically sound and factually grounded justification for the final response. This core signal is supplemented by several auxiliary rewards designed to maintain structural and semantic integrity. By optimizing this reward via GRPO, the model learns to generate responses that are both persuasive, to enhance Satisfaction, and factually accurate, to ensure Faithfulness.

As illustrated in Figure 3, our end-to-end training is driven by the GRPO algorithm. For the t-th turn given input $\mathcal{P} \oplus \mathcal{H}$, the model first performs G parallel rollouts to generate a group of candidate sequences $\{A_t^{(i)}\}_{i=1}^{G}$. The algorithm then uses the reward signal $R^{(i)}$ from each sequence to compute a normalized group advantage score, $\mathcal{A}^{(i)}$, and subsequently updates the policy model.

Reward Function Design The total reward R is a weighted sum of four components, categorized into core and auxiliary rewards, evaluating different aspects of the generated sequence $A_t^{(i)}$ against the ground-truth reference A_t^*:

$$R(A_t^{(i)}, A_t^*) = \sum_{k \in \{\text{bayes, format, len, sem}\}} w_k R_k(A_t^{(i)}, A_t^*) \quad (2)$$

where w_k are hyperparameter weights.

Core: Bayesian-Supervised Reasoning (R_{bayes}). This reward guides the model's internal reasoning chain, Th_t. Grounded in Bayesian principles, our objective is to align this reasoning chain with the reference answer A_t^* by maximizing their joint probability, $P(Th_t, A_t^*)$. Accordingly, the reward is defined as the log-joint probability, which decom-

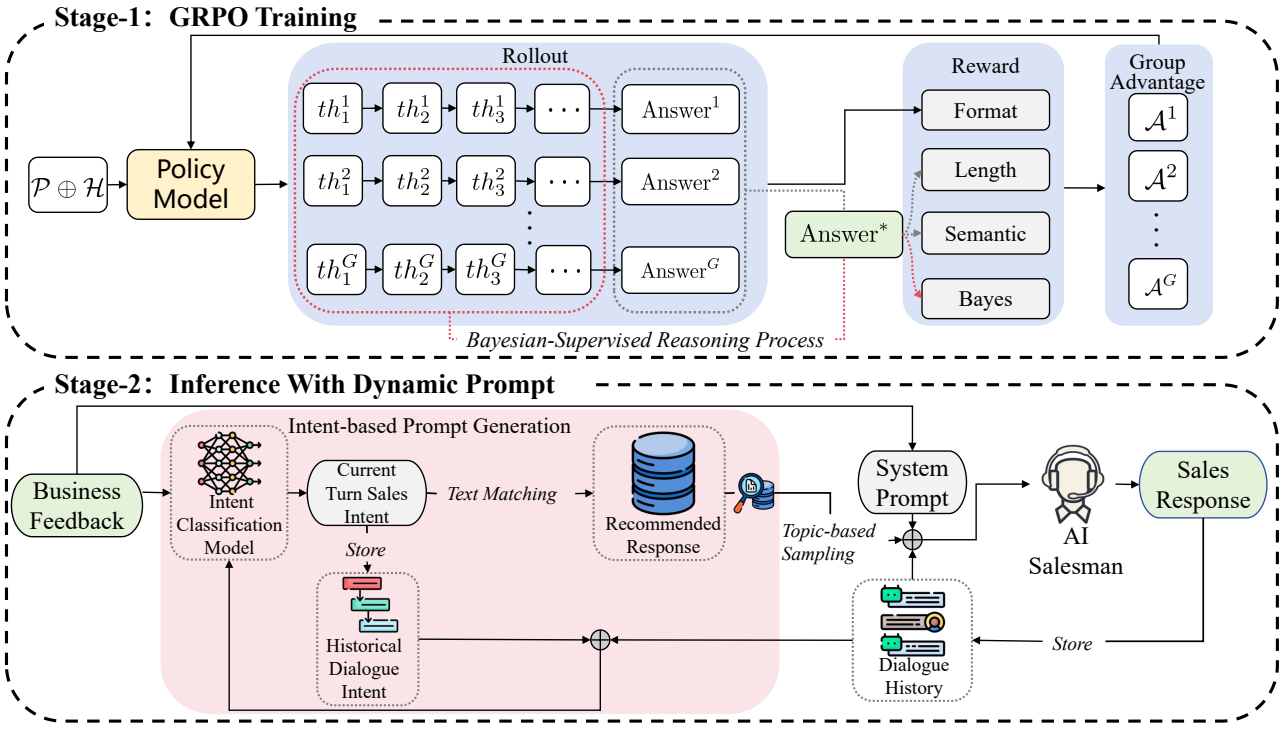


Figure 3: AI Salesman Framework Overview.

poses into two terms estimated by the model π_θ itself.

$$R_{\text{bayes}}(Th_t^{(i)}, A_t^*) = \underbrace{\sum_{j=1}^{m} \log \pi_\theta(th_j^{(i)} \mid \hat{\mathcal{P}}, th_{<j}^{(i)})}_{\text{Prior: Reasoning Fluency}} + \underbrace{\sum_{k=1}^{n} \log \pi_\theta(y_k^* \mid \hat{\mathcal{P}}, Th_t^{(i)}, y_{<k}^*)}_{\text{Likelihood: Reasoning Utility}} \tag{3}$$

where $\hat{\mathcal{P}}$ is the shared context.

Auxiliary 1: Format Adherence (R_{format}). A reward that ensures the output follows the predefined "<think>...</think><answer>...</answer>" schema.

$$R_{\text{format}}(A_t^{(i)}) = f_{\text{format}}(A_t^{(i)}) \tag{4}$$

where $f_{\text{format}}(\cdot)$ is a function that yields 1 if the sequence $A_t^{(i)}$ conforms to the required schema, and 0 otherwise.

Auxiliary 2: Relative Length Consistency (R_{len}). This aims to align the output length with the reference answer by penalizing the squared relative deviation from the target length $L(A_t^*)$.

$$R_{\text{len}}(A_t^{(i)}, A_t^*) = 1 - \left(\frac{|L(A_t^{(i)}) - L(A_t^*)|}{L(A_t^*)} \right)^2 \tag{5}$$

Auxiliary 3: Semantic Similarity (R_{sem}). To measure semantic alignment, we compute the cosine similarity $s^{(i)}$ between the generated and reference answers using a sentence-embedding model. The score is normalized against a baseline similarity s_{base} for a more robust signal.

$$R_{\text{sem}}(A_t^{(i)}, A_t^*) = \frac{s^{(i)} - s_{\text{base}}}{1 - s_{\text{base}} + \epsilon} \tag{6}$$

3.3 Stage-2: Inference With Dynamic Prompt

We propose the Dynamic Outline-Guided Agent (DOGA) to enable customization in telemarketing by overcoming the rigidity of static prompts. Our framework decouples high-level strategy from turn-level execution by generating turn-specific guidance from a pre-structured script library. This process is composed of two stages: an offline library construction phase and a real-time dynamic prompt assembly pipeline. This structure ensures that model responses are personalized and contextually appropriate.

Offline Stage: Structured Script Library Construction The foundation of our framework is a high-quality library of sales scripts and templates. This library is created offline by extracting, clustering, and summarizing effective strategies from a corpus of successful historical dialogues. This process distills best practices into a reusable resource indexed by dialogue intent.

Online Stage: Real-time Dialogue Management During a live conversation, DOGA employs the real-time pipeline shown in Figure 3. At each turn, an Intent Classification

Capability	Model	Mean	Gui.	Fac.	Log.	Use.	Res.	Saf.	Com.
Role-playing	Baseline	5.54	4.83	5.70	5.94	5.40	5.02	7.20	4.68
	SFT-only	5.66	4.78	5.90	6.05	5.61	5.08	7.27	4.92
	GRPO w/ SFT	5.75	4.79	5.95	6.16	5.72	5.08	7.41	5.13
	Ours	**6.31**	**5.81**	**6.5**	**6.62**	**6.16**	**5.76**	**7.6**	**5.75**
Business Analysis	Baseline	6.49	5.42	6.44	7.05	6.67	6.19	7.72	5.91
	SFT-only	6.78	5.39	7.15	7.24	7.04	6.39	7.83	6.41
	GRPO w/ SFT	6.86	5.51	7.39	7.24	6.97	6.59	7.88	6.44
	Ours	**7.40**	**5.96**	**7.87**	**7.78**	**7.61**	**7.23**	**7.94**	**7.43**
Activity Introduction	Baseline	5.91	5.39	5.28	6.43	6.18	5.76	7.39	4.97
	SFT-only	5.86	5.16	5.32	6.38	6.23	5.56	7.33	5.04
	GRPO w/ SFT	5.94	5.08	5.41	6.49	6.15	5.62	7.45	5.36
	Ours	**6.75**	**6.55**	**5.98**	**7.13**	**7.07**	**6.71**	**7.94**	**5.86**
Idle-chat Rejection	Baseline	4.66	4.36	4.35	5.10	4.48	4.41	6.31	3.59
	SFT-only	4.86	3.96	4.68	5.36	4.83	4.67	6.63	3.89
	GRPO w/ SFT	4.95	4.11	4.72	5.48	4.90	4.59	6.78	4.09
	Ours	**5.73**	**5.49**	**5.52**	**6.19**	**5.59**	**5.50**	**6.99**	**4.81**
Objection Handling	Baseline	4.77	4.60	3.92	5.18	4.97	4.47	6.46	3.80
	SFT-only	5.24	5.19	4.64	5.75	5.22	5.01	6.56	4.34
	GRPO w/ SFT	5.33	5.41	4.58	5.82	5.09	5.23	6.69	4.49
	Ours	**6.00**	**6.24**	**4.65**	**6.57**	**6.22**	**6.02**	**7.49**	**4.82**
Operational Guidance	Baseline	5.39	4.44	6.13	5.52	5.33	4.68	6.56	5.09
	SFT-only	5.71	4.84	6.15	5.87	5.54	5.29	6.99	5.29
	GRPO w/ SFT	5.78	4.90	6.20	5.81	5.68	5.16	7.18	5.51
	Ours	**6.74**	**6.26**	**7.33**	**6.71**	**6.50**	**6.09**	**7.63**	**6.67**

Table 1: Performance comparison of different training pipelines. Our framework significantly outperforms all competing baselines. The top-performing model, Ours, utilizes direct reinforcement learning, bypassing the SFT stage. Best results in each block are in **bold**. The second-best results in each block are underlined.

Model first predicts the user's current turn sales intent. This intent is used to retrieve a relevant recommended response template from our pre-built library. Finally, this turn-specific guidance is combined with the system prompt and the full dialogue history to assemble a dynamic system prompt. This prompt steers the model to generate a response that is strategically aligned with the immediate conversational goal.

4 Experiments

This section presents a series of experiments designed to evaluate the effectiveness of our proposed AI-Salesman framework. We first detail the experimental setup, including the datasets, models, and evaluation protocols. We then present the main results comparing our full method against several baselines. Finally, through extensive ablation studies, scalability analysis, and human evaluations, we validate the contributions of the key components of our framework.

Datasets We utilize two datasets with distinct roles in our experiments:

- **TeleSalesCorpus (Syn-Data)**: To ensure the reproducibility and openness of our research, we introduce this synthetic dataset, which will be made publicly available. Constructed as described in Section 3.1, it contains 2,000 high-fidelity, multi-turn dialogues.

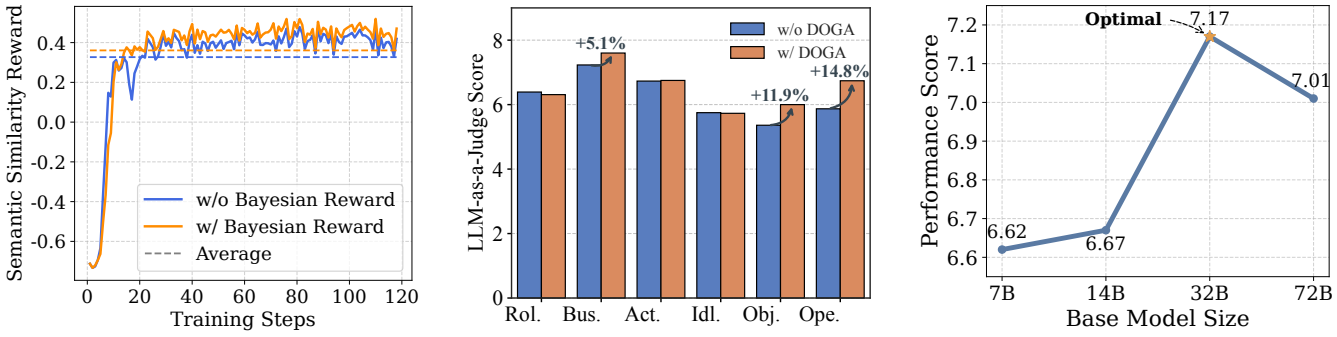
Model	SFT	GRPO	Inference Strategy
Baseline	✗	✗	Few-shot
SFT-only	✓	✗	Few-shot
GRPO w/ SFT	✓	✓	DOGA
Ours	✗	✓	DOGA

Table 2: Configurations for the different models in our experiments. The base model for all versions is **Qwen2.5-7B-Instruct**. ✓ indicates the stage was applied, while ✗ indicates it was skipped.

- **Real-world Tele-sales Dataset (Real-Data)**: This is our dataset for large-scale training. It consists of over 8,000 real-world tele-sales dialogues. This proprietary dataset reflects the complexities of authentic sales conversations, including significant conversational noise and diversity. It is instrumental for assessing our model's performance and scalability in a realistic application setting.

4.1 Experimental Setup

Models and Baselines Our main experiment is based on the Qwen2.5-7B-Instruct model (Qwen et al. 2025). As detailed training and inference configuration in Table 2. We es-



(a) Effect of R_{bayes} on Syn-Data. (b) Advantage in strategic capabilities. (c) Performance scaling with model size.

Figure 4: Key experimental results. (a) Bayesian reward (R_{bayes}) stably raises the upper bound of the semantic similarity reward. (b) DOGA shows decisive advantages in complex, strategic capabilities. (c) Our method's performance scales effectively, with the 32B model offering an optimal trade-off.

tablish a performance reference using the original **Baseline** and a standard Supervised Fine-Tuning **SFT-only model**. Our primary contribution is **Ours**, which applies the GRPO algorithm with the reward function we designed directly to the baseline. To investigate whether SFT is a necessary step for effective preference alignment, we also train a **GRPO w/ SFT** model by applying GRPO after the SFT stage.

Evaluation Metrics As detailed in Section 2.2, we use the LLM-as-a-Judge paradigm with GPT-4 as the evaluator. Each dialogue turn is scored from 1 to 10 across seven metrics. The final score for each of the six core sales capabilities is the arithmetic mean of seven metrics.

4.2 Main Results

The comprehensive performance evaluation, detailed in Table 1, empirically substantiates the remarkable efficacy of our proposed training paradigm. It denoted **Ours**, establishes a new state-of-the-art, achieving dominant scores across the vast majority of capabilities and dimensions evaluated. Our analysis reveals three principal findings:

- **Finding 1: Domain-specific SFT establishes a robust but limited performance baseline.** The results indicate a mixed but overall positive effect from SFT. This confirms its role as a preliminary adaptation stage. While SFT led to significant gains in areas like Business Analysis (6.49 → 6.78) and Objection Handling (4.77 → 5.24), its impact on more complex skills was limited. For example, the score for Role-playing grew minimally from 5.54 to 5.66. This demonstrates that SFT is effective at mimicking explicit patterns but struggles with tasks requiring deeper strategic generalization.
- **Finding 2: SFT creates a performance bottleneck for reinforcement learning.** Our experiments show that applying reinforcement learning to an SFT-initialized model (GRPO w/ SFT) offers negligible performance gain over the SFT model alone, with the overall mean score across all capabilities only increasing minimally from 5.69 (SFT-only) to 5.77 (GRPO w/ SFT). We conclude that SFT, by forcing the model to mimic a noisy and suboptimal dataset, traps its policy in a narrow, flawed space. This severely restricts RL's ability to explore and discover superior strategies, resulting in a final policy that fails to meaningfully diverge from the flawed behaviors learned during SFT. The model thus adheres to rules but lacks conversational richness.
- **Finding 3: Direct RL optimization without SFT unlocks superior performance.** In stark contrast, optimizing a base model directly with our GRPO reward signal yields a holistically superior model, boosting the overall mean score from the Baseline's 5.46 to 6.49—a significant 18.9% increase. By being liberated from the constraints of imitating a potentially suboptimal reference corpus, the model learns to internalize the underlying business logic and knowledge directly from rewards. This approach achieves high performance across all dimensions—excelling not only in Richness (Res.) and User Satisfaction (Use.) but also maintaining strong Guideline Adherence (Gui.), proving it's a more effective path to developing a capable and adaptive sales model.

4.3 Ablation Studies

To evaluate the specific contributions of our proposed components, we conducted a series of ablation studies. These experiments are designed to isolate and quantify the impact of our reward functions and DOGA.

Quantitative Analysis of Reward Components We first investigated the individual importance of the key signals in our composite reward function. To do this, we trained two ablated versions of our model:

- **GRPO w/o R_{bayes}**: The model was trained without the Bayesian-Supervised Reasoning Reward, removing the explicit supervision on the internal thought process.
- **GRPO w/o R_{sem}**: The model was trained without the Semantic Similarity Reward, removing the direct pressure to align the final answer with the expert reference.

As shown in Table 3, the results clearly demonstrate the criticality of both components. Removing the Bayesian reward (R_{bayes}) led to a 5.2% drop in the mean score, while

Model Version	Mean Score
GRPO w/o R_{bayes}	6.15
GRPO w/o R_{sem}	6.39
Ours	**6.49**

Table 3: Ablation study of reward components. The LLM-as-a-Judge calculates the mean score across all evaluation capabilities.

Comparison Pair	Win (%)	Tie (%)	Loss (%)
Ours vs. Baseline	**88.5**	7.2	4.3
Ours vs. SFT-only	75.1	17.6	7.3
SFT-only vs. Baseline	68.7	21.4	9.9

Table 4: A/B test results based on head-to-head human preference evaluations.

removing the semantic reward (R_{sem}) caused a 1.5% decrease. This confirms that both reward signals are essential for guiding the model. R_{sem} directly optimizes for output quality, while R_{bayes} ensures the underlying reasoning is sound, which indirectly but powerfully contributes to the generation of high-quality and reliable responses.

Visualizing the Effect of Bayesian Reward To visualize the effect of our most novel component, the Bayesian reward, we plotted the training-time semantic similarity reward on our synthesized dataset, TeleSalesCorpus (SynData). As shown in Figure 4a, the model trained with R_{bayes} converges to a higher semantic similarity ceiling steadily. This suggests that by penalizing illogical thought processes, the Bayesian reward acts as an internal verifier, preventing the model from exploring ineffective generation paths and steering it more directly toward producing answers that are semantically aligned with expert behavior.

Effectiveness of DOGA A comparative analysis of our DOGA framework against a static prompt on six sales capabilities reveals two key findings (Figure 4b):

- **Finding 1: DOGA excels in complex tasks.** It achieved significant performance gains in Business Analysis (+5.1%), Objection Handling (+11.9%), and Operational Guidance (+14.8%). This performance boost is driven by its ability to dynamically adapt, drawing from a library of expert templates to deliver more detailed and accurate contextual guidance in real-time, surpassing the limitations of static prompts.
- **Finding 2: A trade-off exists between strategic precision and conversational naturalness.**The static prompt performed marginally better in Role-playing and Idle-chat Rejection. DOGA's template injection, while precise, can sound formulaic. For simple tasks, the static prompt's direct rules are more efficient than DOGA's complex retrieval cycle.

4.4 Scalability Analysis

To systematically evaluate the scalability of our proposed method, we conducted a scaling experiment using the Qwen2.5-Instruct series of models, which includes variants with 7B, 14B, 32B, and 72B parameters. Each model was trained and subsequently evaluated on our curated Real-Data set. The results are shown in Figure 4c. We observed a non-linear performance trend with several key findings:

- **Marginal Gain:** Scaling from 7B to 14B yields only a minor improvement.
- **Peak Performance:** The 32B model achieves a significantly higher score of 7.17, marking the peak performance across all tested scales.
- **Diminishing Returns:** Further scaling to 72B leads to a slight performance drop.

These findings indicate that the 32B model offers the optimal capacity for our task, effectively leveraging our proposed frameworks.

4.5 Human Evaluation (A/B Test)

To assess real-world performance, we conducted a blind A/B test with 30 front-line sales professionals. These experts, chosen for their deep understanding of sales strategies and real-world business interactions, role-played as clients and engaged in hundreds of sales conversations with three AI models: our AI-Salesman, a strong SFT-only variant, and a Baseline. They then voted on paired responses, evaluating them on persuasiveness and professionalism.

The results in Table 4 establish a clear performance hierarchy: Ours $\gg$ SFT-only $>$ Baseline. Our full model was preferred in 88.5% of matchups against the baseline and 75.1% against the strong SFT-only model. Notably, this performance ranking aligns with the results from our offline evaluations in Table 1, where GPT-4 served as the judge.

This quantitative strength was echoed in qualitative feedback, where evaluators praised our model for its richer, more varied language and a more natural user experience, confirming its practical value in real-world scenarios.

5 Conclusion

This paper introduces AI-Salesman, an end-to-end framework designed to address the limitations of Large Language Models in professional telemarketing scenarios. Our core innovations include a Bayesian-supervised reinforcement learning algorithm to optimize sales dialogue strategies directly, and the Dynamic Outline-Guided Agent mechanism for flexible, real-time conversation management.

We also constructed and released the first real-world-grounded telemarketing dataset, TeleSalesCorpus, for this task. Extensive automated and human evaluations demonstrate that our approach significantly outperforms baseline models in generating persuasive and business-compliant dialogue.

In summary, this work provides a systematic methodology and practical resources for building more effective and reliable goal-oriented persuasive AI.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was supported by National Key R&D Program of China (2024YFC3308000), the Natural Science Foundation of China (No. 62476265, 62306303, 62506354), the Basic Research Program of ISCAS (Grant No. ISCAS-ZD-202401).

References

Atanasova, P.; Camburu, O.-M.; Lioma, C.; Lukasiewicz, T.; Simonsen, J. G.; and Augenstein, I. 2023. Faithfulness Tests for Natural Language Explanations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 283–294. Toronto, Canada: Association for Computational Linguistics.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *The Twelfth International Conference on Learning Representations*.

Chen, J.; Guan, X.; Yuan, Q.; Mo, G.; Zhou, W.; Lu, Y.; Lin, H.; He, B.; Sun, L.; and Han, X. 2025a. ConsistentChat: Building Skeleton-Guided Consistent Multi-Turn Dialogues for Large Language Models from Scratch. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Chen, Y.; Liu, S.; Lyu, Y.; Zhang, C.; Shi, J.; and Xu, T. 2025b. Xiangqi-R1: Enhancing Spatial Strategic Reasoning in LLMs for Chinese Chess via Reinforcement Learning. arXiv:2507.12215.

Chen, Y.; Lyu, Y.; Liu, S.; Zhang, C.; Lv, J.; and Xu, T. 2025c. Think Wider, Detect Sharper: Reinforced Reference Coverage for Document-Level Self-Contradiction Detection. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 1273–1288. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Friedman, D.; Panigrahi, A.; and Chen, D. 2025. Representing Rule-based Chatbots with Transformers. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3155–3180. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Fu, Y.; Peng, H.; Khot, T.; and Lapata, M. 2023. Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback. arXiv:2305.10142.

Gao, J.; Galley, M.; and Li, L. 2018. Neural Approaches to Conversational AI. In Artzi, Y.; and Eisenstein, J., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2–7. Melbourne, Australia: Association for Computational Linguistics.

Ham, D.; Lee, J.-G.; Jang, Y.; and Kim, K.-E. 2020. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 583–592. Online: Association for Computational Linguistics.

He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2333–2343. Brussels, Belgium: Association for Computational Linguistics.

Hosseini-Asl, E.; McCann, B.; Wu, C.-S.; Yavuz, S.; and Socher, R. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems*, volume 33, 20179–20191.

Lin, L.; Lin, Z.; Zeng, Z.; and Ji, R. 2025. Speculative Decoding Reimagined for Multimodal Large Language Models. arXiv:2505.14260.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. Online: Association for Computational Linguistics.

Pan, M. Z.; Cemri, M.; Agrawal, L. A.; Yang, S.; Chopra, B.; Tiwari, R.; Keutzer, K.; Parameswaran, A.; Ramchandran, K.; Klein, D.; Gonzalez, J. E.; Zaharia, M.; and Stoica, I. 2025. Why Do Multiagent Systems Fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rawte, V.; Chakraborty, S.; Pathak, A.; Sarkar, A.; Tonmoy, S. T. I.; Chadha, A.; Sheth, A.; and Das, A. 2023. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2541–2573. Singapore: Association for Computational Linguistics.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325. Online: Association for Computational Linguistics.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Valmeekam, K.; Marquez, M.; Sreedharan, S.; and Kambhampati, S. 2023. On the Planning Abilities of Large Language Models - A Critical Investigation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5635–5649. Florence, Italy: Association for Computational Linguistics.

Xu, H.-D.; Mao, X.-L.; Yang, P.; Sun, F.; and Huang, H. 2024. Rethinking Task-Oriented Dialogue Systems: From Complex Modularity to Zero-Shot Autonomous Agent. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2748–2763. Bangkok, Thailand: Association for Computational Linguistics.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.