

SPAR: Experience-Driven Strategy Evolution in Multi-Session B2B Sales

Hangming Zhang^{1*†} Yao Zhao^{2*} Haotian Kuang²
C. L. Philip Chen¹ Tong Zhang¹ Tengyue Ma^{2‡}

¹South China University of Technology

²Alibaba Group

hangmingzhang37@gmail.com, philip.chen@ieee.org, tony@scut.edu.cn,
{zhaoyao.zhao, kuanghaotian.kht, xianyue.mty}@taobao.com

Abstract

B2B outbound sales involves high transaction values and relationship-based trust, so closing a deal typically requires multiple rounds of outreach building on prior context. Existing LLM-based sales agents operate within single sessions and cannot carry forward context or refine strategies from experience. Methods that do learn from experience store lessons as unstructured text and treat all outcomes as equally informative, with no mechanism to distinguish a flawed strategy from an external factor. We propose SPAR, a self-evolving framework that models selling strategies as testable hypotheses and refines them without parameter updates. A double-loop architecture lets an inner loop adapt to individual merchants through episodic memory while an outer loop evolves a shared strategy playbook through causal attribution and hypothesis-driven evolution. In a live A/B test on the 1688 platform with over 200,000 merchants, SPAR achieves +15.8% conversion lift and +35.8% high-intent merchant growth. Human evaluation by domain experts rates SPAR at 4.62 out of 5 on overall sales quality, the highest among all methods.

1 Introduction

In B2B outbound sales, high transaction values, multi-stakeholder decisions, and relationship-based trust mean that closing a deal typically requires multiple rounds of persistent, targeted outreach. Experienced salespeople handle this by carrying forward context, judging which approaches work, and refining their playbook across merchants. Large language models have driven interest in automating this process (Chiu et al., 2022; Chang and Chen, 2024; Kim et al., 2025), yet existing systems remain limited to single-session interactions (He

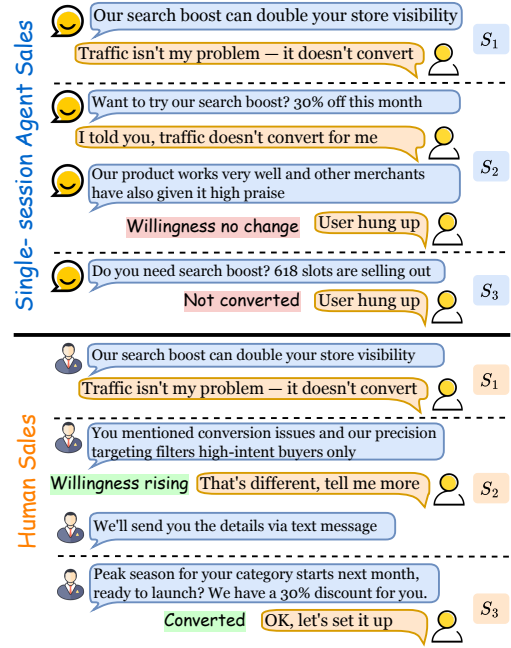


Figure 1: Single-session agents repeat generic pitches across isolated calls without addressing the merchant’s core objection, leading to repeated rejections. Multi-session human salespeople follow up on specific concerns, build on prior context, and achieve conversion. $S_1, S_2, \dots$ denote session 1, session 2, and so on.

et al., 2018; Shea et al., 2024), leaving three closely linked challenges unaddressed.

Current systems treat each session as an isolated event, a **single-session limitation** that caps conversion at first-contact performance (Chang and Chen, 2024), as Figure 1 illustrates. Even with multi-session capability, **attribution ambiguity** persists, as it remains unclear whether a failed follow-up reflects a flawed strategy or an external factor. This ambiguity entrenches **manual dependence** on hand-crafted scripts, since automating strategy updates without reliable attribution would amplify rather than correct errors (Deng et al., 2024). These three barriers operate at successive levels, and each must be resolved for the next to become effective.

While recent LLM-based agents have explored

*Equal contribution.

†Work done during internship at Alibaba Group.

‡Corresponding author.

experience-driven self-improvement (Shinn et al., 2023; Zhao et al., 2024; Park et al., 2023), they treat all outcomes as equally informative and store experience as unstructured text, offering no mechanism to filter external noise or to systematically validate and retire strategies.

We propose SPAR (Strategy Playbook via Attributed Refinement), a self-evolving framework that models selling strategies as *testable hypotheses* and iteratively refines them through sequential merchant interactions without parameter updates. Learning is organized as a double-loop architecture, where an inner loop adapts to individual merchants through multi-session memory, while an outer loop systematically evolves the shared strategy playbook across merchants through causal attribution and hypothesis-driven evolution.

Our contributions are as follows.

- We formulate multi-session B2B sales as a double-loop learning problem and propose SPAR, a self-evolving framework that separates within-merchant adaptation from cross-merchant strategy refinement, requiring no parameter updates.
- We model selling strategies as testable hypotheses with an evidence-based lifecycle of proposal, testing, validation, and retirement, and introduce causal attribution to isolate genuinely strategy-driven outcomes for reliable playbook updates.
- We evaluate SPAR in a live A/B test on the 1688 platform with over 200,000 merchants, achieving +15.8% conversion lift and +35.8% high-intent merchant growth. Simulation analysis also shows SPAR outperforms the strongest baseline by 9.6% on objection recognition and 5.5% on evidence support.

2 Related Work

2.1 Sales Dialogue Agents

Conversational agents for sales and negotiation (Zhan et al., 2024) have progressed from scripted pipelines (Chiu et al., 2022) to LLM-powered approaches that inject strategies through fine-tuning (Chang and Chen, 2024) and learn adaptive policies (Deng et al., 2024). Complementary efforts evaluate negotiation capabilities (Bianchi et al., 2024) and profile users for personalization (Kim et al., 2025; Gou et al., 2025). Despite this progress,

all existing systems operate within a **single interaction session**, with no mechanism to revisit merchants or refine strategies from ongoing experience.

Multi-session dialogue has been studied in open-domain settings (Xu et al., 2022; Maharana et al., 2024; Li et al., 2025). Memory systems and benchmarks built for multi-session agents follow the same framing, where MemMachine (Wang et al., 2026) preserves one user’s conversational record at the episode level for cheap personalized retrieval and MemoryArena (He et al., 2026) tests whether an agent uses what earlier sessions gave it rather than recalling it on demand. This line of work is organized around the record of a single user and improves what an agent can recall, but these systems maintain conversational consistency rather than adapt strategically toward conversion.

2.2 Outcome Attribution

Determining which action caused an observed outcome is fundamental to learning from sequential interactions. Counterfactual credit assignment in reinforcement learning (Mesnard et al., 2021) distinguishes the agent’s contribution from external factors. Despite this importance, LLM agent learning frameworks treat all results as equally informative. SPAR bridges this gap by separating strategy-driven outcomes from externally-determined ones before updating the playbook.

2.3 Experience-Driven Self-Improvement

Building on foundational agent frameworks (Yao et al., 2022), recent work enables LLM agents to learn from experience without parameter updates through self-reflection (Shinn et al., 2023; Madaan et al., 2023; Park et al., 2023), cross-episode transfer (Zhao et al., 2024; Fu et al., 2024; Wang et al., 2024), self-generated examples (Sarukkai et al., 2026), and evolving context (Zhang et al., 2026). However, these methods store experience as unstructured text or flat rules, with no lifecycle to promote validated strategies and retire failing ones.

3 Method

3.1 Task Formulation

The agent contacts N merchants sequentially, each with behavioral profile p_i . For merchant i , the agent conducts up to K dialogue sessions. Session k produces outcome $o_{i,k} \in \{1, 0, -1\}$, representing success, ongoing, and failure respectively. The agent maintains two evolving knowledge stores. An

episodic memory M_i for each merchant records structured session-level notes across follow-ups. A shared **strategy playbook** $\mathcal{P} = \{h_1, \dots, h_m\}$ accumulates selling strategies refined through experience across all merchants. The objective is to maximize the cumulative conversion rate

$$R = \frac{1}{N} \sum_{i=1}^N \max_{k \leq K} \mathbf{1}[o_{i,k} = 1] \quad (1)$$

while simultaneously improving $\mathcal{P}$ such that later merchants benefit from validated lessons learned from earlier ones.

3.2 Framework Overview

SPAR organizes learning as a **double-loop architecture**, illustrated in Figure 2. The **inner loop** adapts to individual merchants across multiple sessions using episodic memory M_i . The **outer loop** evolves $\mathcal{P}$ across merchants by treating each interaction as a natural experiment that tests and refines strategy hypotheses. Algorithm 1 summarizes the complete pipeline, and the following subsections detail each component.

3.3 Strategy as Testable Hypothesis

We model each strategy as a **testable hypothesis** with accumulated evidence that drives lifecycle transitions. The strategy playbook $\mathcal{P}$ is a collection of such hypotheses, each evolved through reinforcement, refinement, addition, merging, and deprecation (§3.5). Each hypothesis $h \in \mathcal{P}$ carries an evidence set $E(h)$, where each entry $e \in E(h)$ records the merchant context and interaction outcome. A causal attribution function $a(e) \in \{s, x, u\}$ then labels the outcome as strategy-driven, externally-determined, or uncertain. Let $E_s(h) = \{e \in E(h) \mid a(e) = s\}$ denote the strategy-attributed evidence and $E_s^{\text{succ}}(h)$ the subset with successful outcomes. The **confidence** of h is computed as the posterior mean of a Beta-Bernoulli model with uniform prior

$$c(h) = \frac{1 + |E_s^{\text{succ}}(h)|}{2 + |E_s(h)|} \quad (2)$$

which starts at $c(h) = 0.5$, avoids overconfidence from small samples, and converges to the empirical success rate as evidence accumulates, isolating the causal effect of the strategy by excluding externally-attributed evidence. Each hypothesis has status $\sigma(h) \in \{0, +1, -1\}$ for tentative, active,

and deprecated, transitioning by

$$\sigma(h) \leftarrow \begin{cases} +1 & |E_s(h)| \geq \tau_p \wedge c(h) \geq \theta_p \\ -1 & |E_s(h)| \geq \tau_d \wedge c(h) < \theta_d \\ \sigma(h) & \text{otherwise} \end{cases} \quad (3)$$

where (τ_p, θ_p) and (τ_d, θ_d) are promotion and deprecation thresholds. $\mathcal{P}^+ = \{h \in \mathcal{P} \mid \sigma(h) \geq 0\}$ denotes the non-deprecated hypotheses. To balance exploitation and exploration, each hypothesis carries an **exploration priority** combining Bernoulli outcome variance with a sample-size exploration bonus

$$I(h) = c(h) \cdot (1 - c(h)) + \frac{1}{1 + n} \quad (4)$$

where $n = |E_s(h)|$.

3.4 Within-Merchant Adaptation

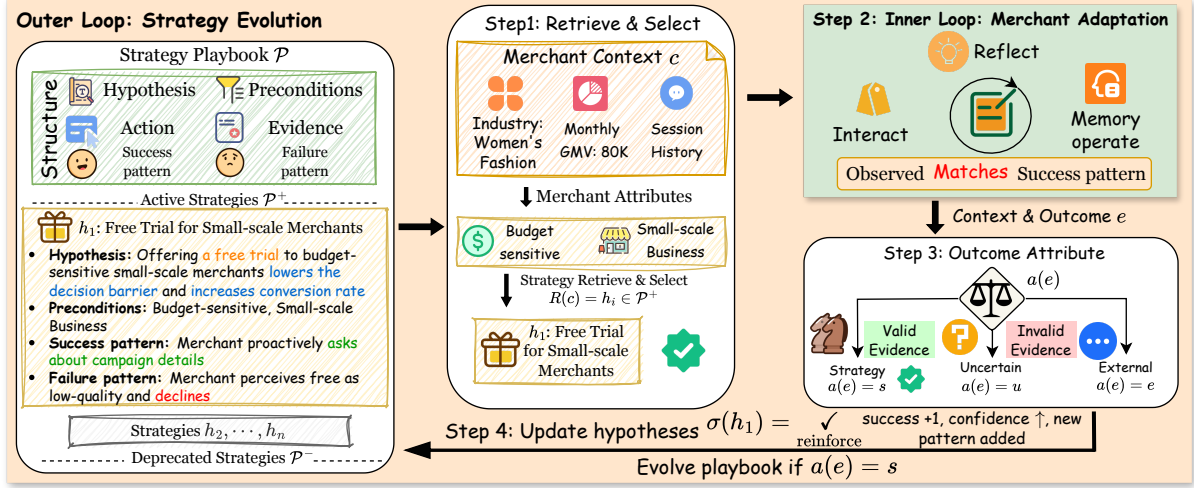
The inner loop maintains an **episodic memory** M_i for each merchant. After each session, the agent generates a structured **reflection** summarizing what happened, why the outcome occurred, and what to try next, then stores it in M_i alongside the outcome and strategy used. This reflection provides actionable context for follow-up sessions and supplies the qualitative signal needed for causal attribution in the outer loop.

At the start of each session, the agent concatenates the merchant profile p_i , interaction history M_i , and the non-deprecated hypothesis set $\mathcal{P}^+$ as selection context. An LLM-based selector implements the scoring function f in Algorithm 1, choosing $h^* \in \mathcal{P}^+$ by evaluating precondition alignment, consistency with prior outcomes, and exploration priority $I(h)$. In follow-up sessions, recorded failure modes steer the selector away from repeating unsuccessful strategies.

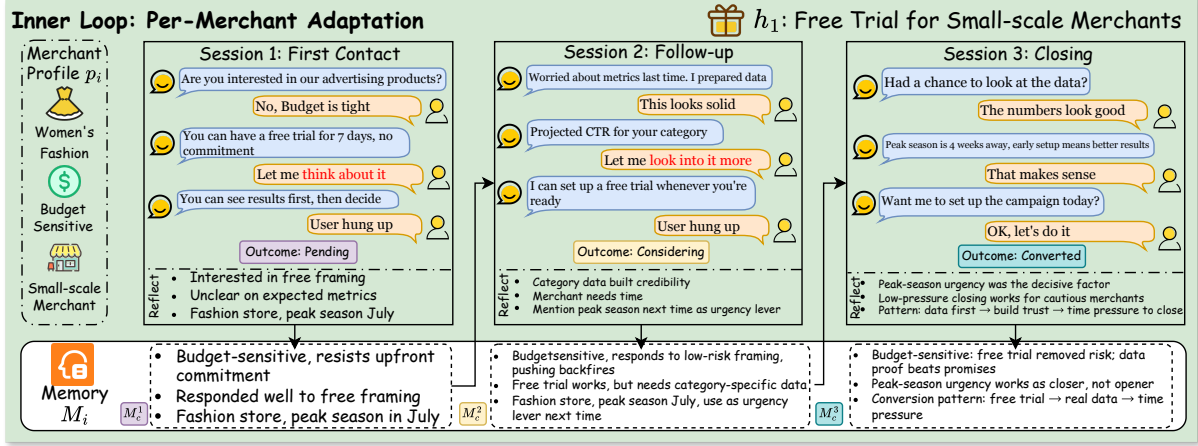
3.5 Cross-Merchant Strategy Evolution

After each merchant interaction, the outer loop transforms raw experience into systematic playbook improvements through causal attribution, hypothesis evolution, and quality curation.

Causal Attribution. The LLM examines the dialogue, merchant profile, outcome, and strategy used, then assigns an attribution label $a(e) \in \{s, x, u\}$. This counterfactual credit assignment (Mesnard et al., 2021) gates playbook updates. Only outcomes with $a(e) = s$ enter $E_s(h^*)$ and trigger evolution. External outcomes are recorded but do not affect $c(h)$.



(a) The double-loop architecture. The outer loop maintains and evolves a strategy playbook $\mathcal{P}$ of testable hypotheses through four steps after each interaction. The inner loop adapts to individual merchants across sessions using episodic memory M_i .



(b) Detailed view of the inner loop for a single merchant across three sessions. Episodic memory accumulates observations from each session and guides follow-up strategies, enabling the agent to build rapport and adapt its approach over time.

Figure 2: Overview of the SPAR framework. All strategies and merchant data shown are anonymized. The illustration is a simplified example for readability. Real B2B sales scenarios involve more complex merchant profiles and longer interaction sequences.

Hypothesis Evolution. For outcomes causally attributed to the strategy, the agent applies structured delta updates to the relevant hypothesis

$$\mathcal{P} \leftarrow \Phi(\mathcal{P}, h^*, e) \quad (5)$$

where e is the newly constructed evidence. The LLM implements Φ by selecting from seven evolution actions, designed around a **minimal revision principle** that always prefers the smallest targeted change sufficient to incorporate new evidence. *Reinforce*, *Update*, *Condition*, and *None* preserve existing hypotheses with increasing degrees of modification. *Add*, *Merge*, and *Deprecate* restructure the playbook more aggressively when incremental updates are insufficient. Together these complementary actions cover evidence accumulation, hypothesis refinement, and structural reorganization.

Exploration and Curation. The agent performs **fork discovery** after each interaction, identifying unexplored strategy directions from the dialogue as exploration candidates. The playbook is deduplicated at regular intervals, and the deprecation rule in Eq. 3 is applied after every interaction.

4 Experiments

4.1 Experimental Setup

Dataset. We construct a multi-session B2B sales evaluation dataset from 1,642 real conversations between an experienced human sales team and merchants on 1688, a major B2B online platform. Each conversation is manually annotated with detailed merchant profiles, behavioral archetypes, session-level outcomes, and strategy labels.

Metric	Score
Outcome agreement	98.7%
Key-behavior recall	91%
Turn-count deviation	0.9 ± 1.3
Paired comparison accuracy	87%
Expert approval	89%

Table 1: Merchant simulator fidelity validated against 300 real conversations and human evaluation from the sales team.

Merchant Simulator. We use GPT-4o as the simulated merchant, conditioned on a detailed profile with target prices, behavioral traits, and session-specific goals. We validate fidelity by replaying 300 real conversations and conducting a human evaluation with the sales team. The simulator is conditioned on the full conversation context from real human merchant interactions, yielding 98.7% outcome agreement as shown in Table 1. The 87% paired comparison accuracy further confirms that the simulator responds differently to strong and weak strategies. Expert approval reaches 89%, validating its use as a reliable evaluation proxy.

Baselines. We compare SPAR against three baselines. **Direct Prompting** uses the same backbone LLM with only the item description and no memory or evolving playbook. **Reflexion** (Shinn et al., 2023) generates a natural-language lesson after each conversation and stores it per merchant as self-reflective feedback for future interactions. **ExpeL** (Zhao et al., 2024) extracts reusable insights from past interactions and retrieves them based on task similarity to guide decision-making.

Metrics. We evaluate agents along three dimensions of sales dialogue quality.

- **Objection handling** measures whether the agent recognizes the merchant’s core objection, responds to it directly, and supports its response with concrete evidence.
- **Dialogue quality** is assessed through *monologue rate*, the fraction of sessions dominated by lengthy seller utterances, and *repetition rate*, the fraction of follow-up sessions that duplicate earlier questions or arguments.
- **Memory recall** measures whether the agent proactively references specific facts from prior sessions before the merchant raises them, evaluated only after the first session.

For LLM-based evaluation, all metrics except repetition rate are scored by an LLM evaluator (Zheng

et al., 2023) at the session level across all real conversations. Repetition rate is computed via text similarity between consecutive sessions.

Hypothesis Design. The hypothesis schema is derived from LLM-assisted analysis of 51,603 session notes from a top-performing B2B sales team on 1688 and validated against 3,637 sales cases. Each hypothesis contains six fields: *description*, *preconditions*, *action steps*, *rationale*, *success and failure patterns*, and an *evidence* set with causal attribution labels. Confidence scoring and lifecycle thresholds are similarly grounded in observed patterns of how experienced salespeople adopt and retire strategies. The session notes used for schema design have no overlap with the 1,642 evaluation conversations.

Implementation. All agents use GPT-4o as the backbone LLM. The playbook is initialized with a small set of domain-specific tentative seed strategies and capped at 50 entries.

4.2 Results and Ablation

LLM-based Evaluation. Table 2 presents results across baselines and ablation variants under two judge models, where SPAR **leads on most metrics** under both. Under GPT Eval, the advantage is most pronounced on **evidence support**, where SPAR achieves **65.6% higher** than Direct Prompting and **5.5% higher** than the strongest baseline, showing that the double-loop architecture yields more substantive selling arguments. SPAR also achieves the lowest repetition rate at 0.002 and the highest memory recall at 0.071. ExpeL and Reflexion improve through experience accumulation, yet both lack the structured hypothesis lifecycle that allows SPAR to validate effective strategies and retire failing ones. The recognition gain over the strongest baseline is **9.6%**. Under Claude Eval, SPAR leads the strongest baseline by **16.6%** on response, cuts the repetition rate to **0.003** from 0.006, and raises memory recall to **0.064** from 0.042.

Human Evaluation. Three domain experts annotate dialogues with 300+ real merchants under blind conditions (Table 2). SPAR achieves the **highest scores across all seven metrics**, with a Sales Quality of **4.62 out of 5** and evidence support at **0.926**. The advantage is most visible on **repetition** at **0.132**, an **18.5%** reduction over the next-best method, confirming that playbook evolution produces non-repetitive follow-up interactions. Di-

Method	Objection Handling			Dialogue Quality		Memory	Sales Quality
	Recognize $\uparrow$	Respond $\uparrow$	Evidence $\uparrow$	Monologue $\downarrow$	Repetition $\downarrow$	Recall $\uparrow$	Score $\uparrow$
GPT Eval	Direct Prompting	0.741	0.540	0.428	0.186	0.005	—
	ExpeL	0.705	0.623	0.607	0.262	0.016	—
	Reflexion	0.759	0.638	0.672	0.221	0.017	0.053
	SPAR w/o Inner Loop	0.684	0.579	0.596	0.281	0.018	—
	SPAR w/o Outer Loop	0.800	0.672	0.672	0.272	0.009	0.065
	SPAR w/o Evolution	0.695	0.661	0.627	0.288	0.007	—
	SPAR	0.832	0.729	0.709	0.221	0.002	0.071
Claude Eval	Direct Prompting	0.714	0.492	0.415	0.178	0.006	—
	ExpeL	0.642	0.578	0.548	0.275	0.020	—
	Reflexion	0.728	0.584	0.621	0.211	0.024	0.042
	SPAR w/o Inner Loop	0.635	0.542	0.536	0.289	0.018	—
	SPAR w/o Outer Loop	0.761	0.618	0.628	0.252	0.013	0.059
	SPAR w/o Evolution	0.651	0.622	0.574	0.284	0.009	—
	SPAR	0.798	0.681	0.662	0.205	0.003	0.064
Human Eval	Direct Prompting	0.762	0.884	0.741	0.306	0.346	—
	ExpeL	0.871	0.857	0.879	0.384	0.337	—
	Reflexion	0.846	0.903	0.896	0.318	0.162	0.762
	SPAR	0.921	0.938	0.926	0.258	0.132	0.813
							4.62

Table 2: Results on the B2B sales evaluation dataset. **Bold with underline** marks the best and underline marks the second-best in each column. Recognize, Respond, and Evidence measure objection handling, Monologue and Repetition measure dialogue quality, and Recall measures proactive memory callback in follow-up sessions. GPT Eval and Claude Eval score the same dialogues with GPT-4o and with Claude Sonnet 4.5. Sales Quality is rated by three domain experts on a 1–5 scale under blind conditions (rubric in Appendix C).

rect Prompting scores competitively on response at **0.884**, reflecting the strong zero-shot capability of the backbone LLM, but falls behind on recognition at **0.762** and evidence at **0.741**, where structured learning provides the largest gains. The performance gaps are narrower than in LLM-based evaluation under both judge models, yet the **ranking across methods remains consistent**, confirming that the automated metrics reliably capture quality differences perceived by experienced salespeople.

Ablation Study. The ablation variants isolate each component’s contribution. Under GPT Eval, removing the inner loop causes the **largest degradation**, with recognition dropping **17.8%** and memory recall falling to zero, showing that **per-merchant memory is essential** for cross-session continuity. Removing the outer loop retains strong recognition at only 3.8% below SPAR but lags **7.8%** on response and **5.2%** on evidence, confirming that **the evolving playbook drives deeper objection handling**. Removing evolution freezes the playbook after seeding, falling **9.3%** behind on response and **11.6%** on evidence, as the agent cannot refine strategies from new experience. Under Claude Eval, recognition falls **20.4%** without the inner loop and evidence falls **13.3%** without evolution. These results confirm that **all three components are necessary and mutually reinforcing**.

Method	Conv. $\uparrow$	Sessions $\downarrow$	M1 Rate $\uparrow$
Direct Prompting	—	—	—
ExpeL	+1.1%	+0.3%	+9.2%
Reflexion	+9.6%	−4.6%	+27.9%
SPAR w/o Inner Loop	+10.1%	−6.3%	+34.5%
SPAR w/o Outer Loop	+12.6%	−11.2%	+22.3%
SPAR w/o Evolution	+9.6%	−3.0%	+18.7%
SPAR	+15.8%	−9.2%	+35.8%

Table 3: Online A/B test results on the 1688 platform. All values are relative changes over the Direct Prompting baseline. Conv. is conversion rate, Sessions is the number of outreach sessions needed, and M1 Rate is first-milestone achievement rate. **Bold with underline** marks the best and underline marks the second-best.

4.3 Online A/B Test

To validate that simulation improvements transfer to the real world, we deploy all methods on the 1688 platform in a live A/B test with over 30,000 merchants per group. Conversion is defined as completing a target transaction within 7 days of the final outreach session. The M1 high-intent rate measures how often merchants reach high-intent status during the sales process. Table 3 reports the relative change over the Direct Prompting baseline on three business metrics. Statistical significance is assessed via a two-proportion z-test with 95% CIs from 1,000-iteration bootstrap, and all SPAR lifts are significant at $p < 0.001$.

SPAR achieves the **highest conversion lift at +15.8%** and the **highest M1 achievement lift at +35.8%**, while reducing outreach sessions by **9.2%**. The conversion advantage over the next-best variant is **25.4%**, confirming that the full double-loop architecture delivers the strongest end-to-end selling capability in production. At scale, this yields operational savings while easing merchant fatigue from repeated contact.

The ablation pattern mirrors the simulation results. SPAR w/o Outer Loop achieves the second-highest conversion at +12.6% and the largest session reduction at −11.2%, while SPAR w/o Inner Loop yields +10.1% conversion with the highest M1 lift among ablation variants at +34.5%. SPAR w/o Evolution drops M1 to +18.7%, confirming that **strategy evolution nearly doubles the M1 gain**. ExpeL shows minimal improvement at +1.1% conversion, suggesting that unstructured insight accumulation provides limited benefit in multi-session B2B sales. We provide a qualitative analysis of playbook evolution in Appendix A.

5 Conclusion

We presented SPAR, a self-evolving framework that models selling strategies as testable hypotheses and refines them through a double-loop architecture with causal attribution. In a live A/B test on the 1688 platform with over 200,000 merchants, SPAR achieves +15.8% conversion lift and +35.8% high-intent merchant growth. Simulation analysis shows SPAR outperforms the strongest baseline by 9.6% on objection recognition and 5.5% on evidence support. These results are validated in real-world B2B commercial outreach where merchants make purchasing decisions.

Limitations

SPAR requires multiple interactions before the playbook meaningfully evolves, so early merchants do not benefit from accumulated experience. This cold-start cost is inherent to any online learning approach, and investigating warm-start techniques such as transferring playbooks across domains is a promising direction for future work.

The strategy evolution process relies on LLM reasoning, which inherently introduces a degree of randomness and unpredictability in playbook updates. While the confidence mechanism and lifecycle thresholds provide gradual self-correction over time, individual evolution decisions may be

suboptimal. Incorporating an auxiliary verification agent to validate proposed strategy changes before they enter the playbook is a promising direction for improving reliability.

Merchants are processed sequentially in our current experiments. In practice, sales teams handle multiple merchants concurrently, where parallel interactions may produce fundamentally different learning dynamics. Evaluating SPAR in concurrent settings remains future work.

Acknowledgments

This work is supported by the CCF-1688 Yuanbao Cooperation Fund (No. CCF-Alibaba2025007) and the Key-Area Research and Development Program of Guangdong Province (No. 2023B0303030001).

References

- Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. In *International Conference on Machine Learning*, pages 3935–3951. PMLR.
- Wen Chang and Yun-Nung Chen. 2024. Injecting salesperson’s dialogue strategies in large language models with chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3798–3812.
- Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. Salesbot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158.
- Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *International Conference on Learning Representations*, volume 2024, pages 10058–10090.
- Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. 2024. Autoguide: Automated generation and selection of context-aware guidelines for large language model agents. *Advances in Neural Information Processing Systems*, 37:119919–119948.
- Qi Gou, Zehua Xia, Li Juan, Qingyang Zhao, and Wenjing Yang. 2025. Advancing e-commerce merchants telemarketing with synthetic data-driven llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2146–2154.
- He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in

- negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.
- Zexue He, Jiaxin Pei, and Alex Pentland. 2026. [Memo-ryarena: Benchmarking agent memory in interdependent multi-session agentic tasks](#). In *4th Workshop on Towards Knowledgeable Foundation Models at ACL 2026*.
- Tongyoung Kim, Jeongeun Lee, SooJin Yoon, SungHwan Kim, and Dongha Lee. 2025. [Towards personalized conversational sales agents: Contextual user profiling for strategic action](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5131–5154, Suzhou, China. Association for Computational Linguistics.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! llm-powered personalized agent for long-term dialogue. In *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)*, pages 5259–5276.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.
- Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S Stepleton, Nicolas Heess, Arthur Guez, et al. 2021. Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning*, pages 7654–7664. PMLR.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Vishnu Sarukkai, Zhiqiang Xie, and Kayvon Fatahalian. 2026. Self-generated in-context examples improve llm agents for sequential decision-making tasks. *Advances in Neural Information Processing Systems*, 38:64392–64425.
- Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W Morris, and Zhou Yu. 2024. Ace: A llm-based negotiation coaching system. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 12720–12749.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. [Voyager: An open-ended embodied agent with large language models](#). *Transactions on Machine Learning Research*.
- Shu Wang, Edwin Yu, Oscar Love, Tom Zhang, Tom Wong, Steve Scargall, and Charles Fan. 2026. [Memmachine: A ground-truth-preserving memory system for personalized ai agents](#). *Preprint*, arXiv:2604.04853.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 5180–5197.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Haolan Zhan, Yufei Wang, Zhuang Li, Tao Feng, Yuncheng Hua, Suraj Sharma, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Reza Haf. 2024. Let’s negotiate! a survey of negotiation dialogue systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2019–2031.
- Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. 2026. [Agentic context engineering: Evolving contexts for self-improving language models](#). In *The Fourteenth International Conference on Learning Representations*.
- Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Playbook Evolution Analysis

We examine how the strategy playbook evolves during the online A/B test to illustrate the hypothesis lifecycle. The playbook is initialized with six tentative seed strategies. As larger-scale manual analysis is infeasible, we qualitatively examine its evolution on a sample of 200+ real merchants drawn from the A/B test.

Lifecycle Progression. Two of the six seed strategies are promoted to *active* status through accumulated evidence. Strategy H002, which instructs the agent to reference prior conversations with returning merchants, accumulates 108 strategy-attributed successes and is promoted after reaching the evidence threshold. Strategy H001, which recommends offering trial programs when merchants hesitate, follows a similar trajectory with 89 attributed successes. Two other seed strategies are deprecated after repeated failures, illustrating that the lifecycle correctly retires ineffective approaches. The evolution process also discovers and validates five new strategies beyond the initial seeds, demonstrating that the playbook grows organically from interaction experience.

Fork Discovery. The evolution process identifies 9 unexplored strategy directions from dialogue analysis. These directions capture recurring patterns that existing strategies do not address, such as offering hands-on onboarding support and providing industry-specific case studies to build credibility. Each direction is logged with the number of times it appears across interactions, creating a prioritized queue for future hypothesis generation.

Attribution Filtering. The causal attribution mechanism labels a subset of outcomes as externally driven, preventing them from distorting strategy confidence scores. When a merchant walks away due to a recently reduced internal budget rather than a flaw in the agent’s approach, the attribution module assigns label x and the corresponding evidence does not affect $c(h)$. This filtering ensures that the playbook reflects genuine strategy effectiveness rather than environmental noise.

B Algorithm

Algorithm 1 describes the complete SPAR pipeline. The outer loop iterates over merchants sequentially. For each merchant, the inner loop runs up to K sessions. At each session, the agent selects a strat-

egy hypothesis h^* from the non-deprecated set $\mathcal{P}^+$ using a scoring function that considers precondition alignment, prior interaction history, and exploration priority $I(h)$. After dialogue execution, a reflection is stored in episodic memory M_i . The outer loop then constructs evidence from the interaction, assigns a causal attribution label, and triggers playbook evolution Φ only for strategy-attributed outcomes. The session loop terminates early on deal closure ($o = 1$) or merchant decline with no follow-up opportunity ($o = -1$).

Algorithm 1 SPAR Double-Loop Learning

Require: Merchants $1, \dots, N$ with profiles $\{p_i\}$,
max sessions K , playbook $\mathcal{P}$

Ensure: Evolved playbook $\mathcal{P}$

```

1: for  $i = 1$  to  $N$  do
2:   Initialize episodic memory  $M_i \leftarrow \emptyset$ 
3:   for  $k = 1$  to  $K$  do
4:     // Inner Loop - Strategy Selection
5:      $h^* \leftarrow \arg \max_{h \in \mathcal{P}^+} f(h, p_i, M_i, I(h))$ 
6:     Conduct dialogue using  $h^*$ , observe  $o_{i,k}$ 
7:     Generate reflection, store in  $M_i$ 
8:     // Outer Loop - Evolution
9:     Construct evidence  $e$  from  $p_i, h^*, o_{i,k}$ ,
       and dialogue
10:    Assign attribution  $a(e) \in \{s, x, u\}$ 
11:    Append  $e$  to evidence set  $E(h^*)$ 
12:    if  $a(e) = s$  then
13:       $\mathcal{P} \leftarrow \Phi(\mathcal{P}, h^*, e)$ 
14:    end if
15:    Update  $\sigma(h)$  for all  $h \in \mathcal{P}$ 
16:    if  $o_{i,k} = 1$  then
17:      break // deal closed
18:    else if  $o_{i,k} = -1$  then
19:      break // declined, no follow-up
20:    end if
21:  end for
22: end for
23: return  $\mathcal{P}$ 

```

C Sales Quality Rubric

Three domain experts rate each dialogue on a 1–5 scale based on the following rubric. Each level describes both overall dialogue quality and selling-strategy execution along four dimensions, namely *strategic fit*, *objection handling*, *rapport building*, and *closing effectiveness*.

- **5 – Excellent.** The agent demonstrates outstanding professional handling and effectively guides toward conversion. Strategic fit

matches the merchant's profile and stage precisely, objections are addressed with concrete evidence, rapport is consistently built across sessions, and closing is clear and well-timed. Suitable as a reference example for sales training.

- **4 – Good.** The agent handles the conversation professionally. Strategic fit is generally appropriate, objections are addressed with relevant supporting points, rapport is established across sessions, and closing is clear, with only minor opportunities for improvement.
- **3 – Acceptable.** The agent completes the basic sales flow. Strategic fit is partial, objections are acknowledged but not always supported with concrete evidence, rapport is limited, and closing lacks decisiveness.
- **2 – Poor.** The agent makes noticeable errors. Strategic fit is weak or mismatched, key objections are missed or inadequately handled, rapport is minimal, and no clear closing is offered, reducing the likelihood of conversion.
- **1 – Unacceptable.** The agent provides irrelevant or inaccurate responses. Strategic fit is broken, objections are ignored or aggravated, no rapport is established, and the conversation may harm the merchant relationship.