Sell More, Play Less: Benchmarking LLM Realistic Selling Skill

Anonymous ACL submission

Abstract

Sales dialogues require multi-turn, goal-directed persuasion under asymmetric incentives, which makes them a challenging setting for large language models (LLMs). Yet existing dialogue benchmarks rarely measure deal progression and outcomes. We introduce SalesLLM benchmark, a bilingual (ZH/EN) benchmark derived from realistic applications covering Financial Services and Consumer Goods, built from 30,074 scripted configurations and 1,805 curated multi-turn scenarios with controllable difficulty and personas. We propose a fully automatic evaluation pipeline that combines (i) an LLM judge for sales-process progress, and (ii) fine-tuned BERT classifiers for end-of-dialogue buying intent. To improve simulation fidelity, we train a user model, CustomerLM, with SFT and DPO on 8,000+ crowdworker-involved sales conversations, reducing role inversion from 17.44% (GPT-4o) to 8.8%. SalesLLM benchmark scores correlate strongly with expert human ratings (average Pearson $r = 0.86$ across four models, with 500 human annotations per model). Experiments across 15 mainstream LLMs reveal substantial variability: top-performance LLMs are competitive with human-level performance while the less capable ones are worse than human. SalesLLM benchmark serves as a scalable benchmark for developing and evaluating outcome-oriented sales agents. Our code and data are released at https://anonymous.4open.science/r/SaleLLM-217A/.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in goal-directed interactions, with sales emerging as a high-impact application (Chang and Chen, 2024a; Murakhovs'ka et al., 2023; Chang and Chen, 2024b; Zhang et al., 2026; Cheng et al., 2025). However, existing benchmarks emphasize dialog quality (Yao et al., 2024; Liu et al., 2025, 2016; Sai et al., 2020) rather than outcome-oriented abilities such as proactive influence and conversion, which are the core skills of sales. To be specific, sales represent a distinct form of **asymmetric persuasion**, where agents must influence user attitudes despite potential resistance (Petty and Cacioppo, 1986; Wang et al., 2020). To address this gap, we introduce **SalesLLM benchmark**, a benchmark for evaluating proactive persuasion in realistic sales contexts. We synthesize 30,074 task scripts and curate 1,805 diverse multi-turn scenarios across Financial Services and Consumer Goods. As illustrated in Figure 1, SalesLLM benchmark employs a dual-assessment framework: LLM judges evaluate sales process efficiency, while BERT-based models estimate purchase intention.

To achieve realistic simulation, we build **CustomerLM**, a specialized user simulator trained on 8,000+ crowdworker dialogues using SFT and DPO (Ouyang et al., 2022; Rafailov et al., 2023), addressing formal language bias and role confusion in general-purpose simulators.

Our contributions include: (1) SalesLLM benchmark, a benchmark with 1,805 multi-turn scenarios in Chinese and English; (2) CustomerLM, a realistic user simulator reducing role inversion; (3) an automated dual-scoring framework for sales evaluation; (4) comprehensive evaluation revealing performance gaps in asymmetric persuasion.

2 Related Work

Multi-turn benchmarks like Sotopia (Zhou et al., 2024), DailyPersuasion (Jin et al., 2024), and personalized persuasive dialogue systems (Wang et al., 2019) focus on social or persuasive dialogues but lack explicit conversion objectives and measurable behavioral outcomes central to sales. Existing sales-focused LLM work (Chang and Chen, 2024a; Murakhovs'ka et al., 2023) relies on synthetic or simulated dialogues without systematic evaluation frameworks, while others (Zhang et al., 2026) are

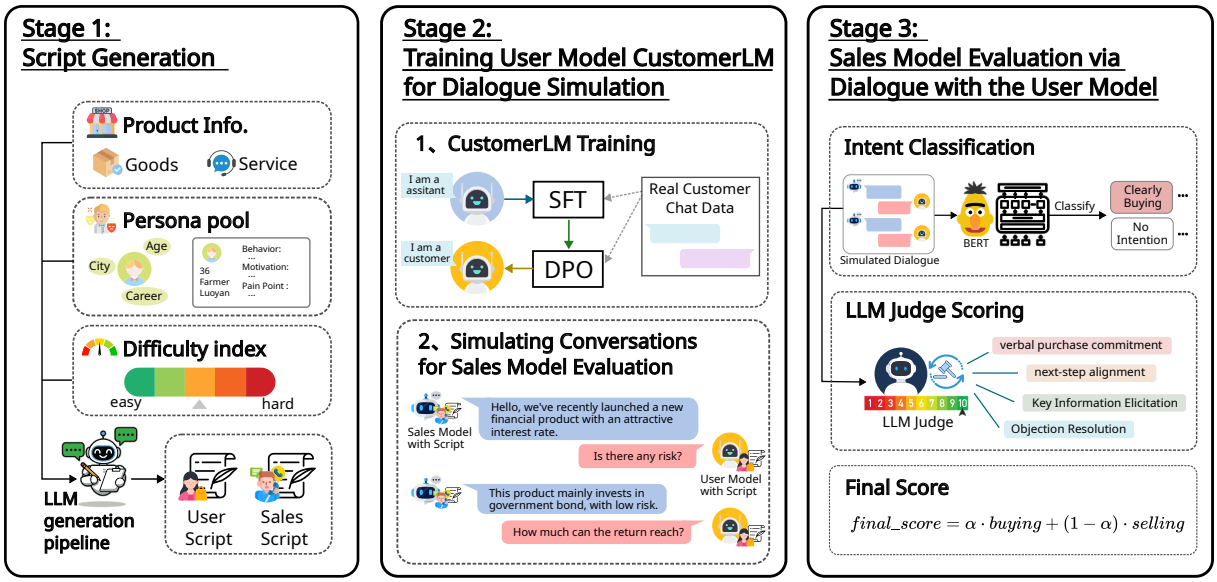


Figure 1: Overview of the SalesLLM benchmark: (1) Standardized Script Generation: Synthesizing role-play scripts by integrating product inventories, persona pools, and difficulty controls. (2) CustomerLM Development & Simulation: Training a specialized simulator via SFT and DPO to conduct simulated dialogues with evaluated sales LLMs. (3) Evaluation Pipeline: Evaluating sales performance through a dual-metric evaluation framework combining BERT-based intent classification for outcomes and LLM-as-a-judge scoring for process quality.

restricted to narrow telemarketing domains without real-world product diversity.

Current LLM-based user simulators produce overly formal responses and role confusion (Wang et al., 2025; Sekulic et al., 2024), limiting evaluation fidelity in outcome-driven interactions. SalesLLM benchmark addresses these gaps with a systematic benchmark for asymmetric persuasion contexts.

3 SalesLLM Benchmark

To evaluate the selling and persuasive capabilities of LLMs, SalesLLM benchmark integrates three core components (Figure 1). First, standardized sales scripts define specific sales scenarios for both salesperson and customer roles. Second, the specialized user simulator, CustomerLM, acts as a virtual customer to engage the evaluated sales LLMs in multi-turn dialogues. Finally, a systematic evaluation pipeline quantifies the overall selling proficiency of the evaluated sales LLMs. We detail the construction of these components in the subsequent sections.

3.1 Script Generation

To evaluate the performance of sales LLMs, we first establish high-fidelity sales scenarios. A realistic scenario integrates a salesperson equipped with comprehensive product knowledge and strategic selling points, alongside a customer characterized by genuine needs, distinct personality traits, and specific purchase propensities. In alignment with established methodologies (Wang et al., 2024b; Ma et al., 2024), we construct our scenario scripts based on three fundamental pillars: Product Inventories, Base Personas, and Controllable Difficulty. Based on these pillars, we synthesize concrete sales scenarios that encompass detailed scripts for both the salesperson and the customer. Figure 2 visualizes this process.

Product Inventories. Our product inventory comprises two primary domains: consumer goods and financial services. For consumer goods, we leverage the Amazon-Reviews-2023 dataset (Hou et al., 2024) as seed information. Specifically, we sample five product descriptions from the same category as context to guide an LLM in synthesizing unique product profiles that include basic specifications, key features, and strategic selling points. Following raw generation, we apply MinHash-based (Broder, 1997) deduplication to ensure diversity, resulting in 10,074 unique consumer products. For financial services, we collect 300 real-world products and employ an identical pipeline to maintain cross-domain consistency, yielding a combined total of 20,000 samples.

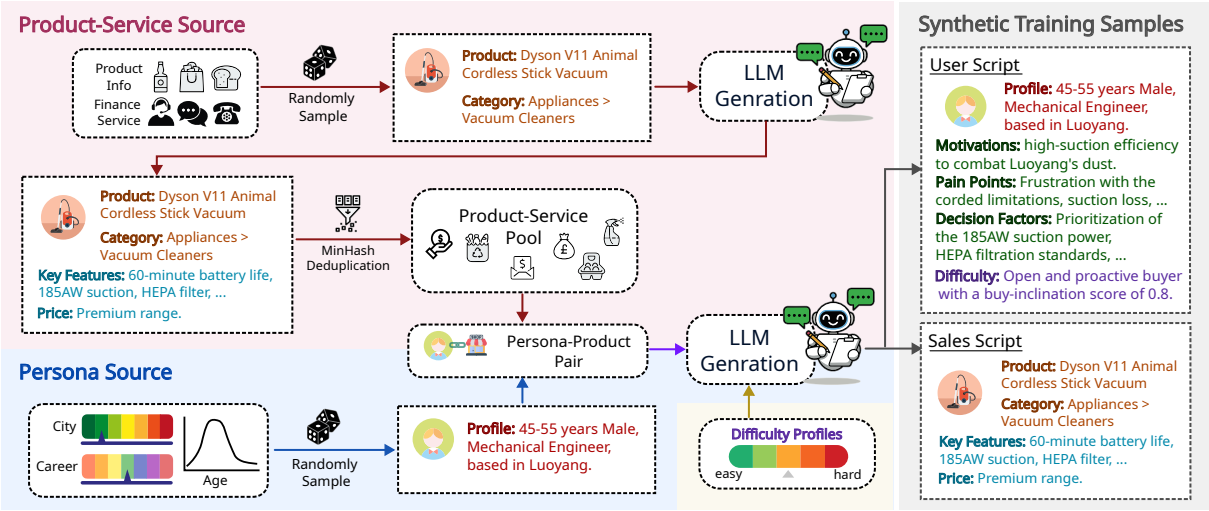


Figure 2: The Workflow of the Script Generation Pipeline. The construction process follows a logical transition from raw resources to synthesized scripts: (1) Raw Component Generation: The pipeline establishes three pillars: Product Information, where raw items are expanded into detailed profiles and deduplicated via MinHash; Persona Generation, involving the sampling of demographic attributes (Age, City, and Career); and the specification of Difficulty Profiles. (2) Persona-Product Pairing: The processed product-service pool is mapped to the sampled personas to form unique Persona-Product Pairs as the contextual foundation for the scenario. (3) LLM-based Information Synthesis: Based on these pairs and the assigned difficulty, LLMs generate the final scripts, including a User Script with product-specific motivations and pain points, alongside a Sales Script containing structured product features and selling points.

Base Personas. We define the foundation of our customer profiles using three core demographic attributes: age, occupation, and city, where the city attribute is integrated with specific decision factors as detailed in Appendix B.2. By performing random sampling across these attributes, we construct a comprehensive pool of base personas. This structured approach provides a diverse population of potential customers, serving as the basis for the more complex, product-conditioned psychological enrichment in the subsequent stage. The distribution of decision factors across different age groups is illustrated in Figure H.1(see Appendix H).

Controllable difficulty and intent. We modulate scenario difficulty, which represents the relative challenge a sales agent faces in successfully promoting a product, through systematic prompt engineering. Specifically, we define five calibrated difficulty tiers: *easy*, *medium*, *hard*, *very hard*, and *adversarial*. Each tier corresponds to a unique system prompt that explicitly incorporates a specific buy propensity and a canonical buyer style (ranging from *cooperative* to *adversarial*); detailed difficulty profiles are provided in Appendix B.3. This controllable design allows the same product and base persona to be instantiated under systematically varied conditions, consistent with behavioral test suites (Ribeiro et al., 2020), user-simulation-based evaluation in task-oriented dialogue (Sun et al., 2024; Luo et al., 2024), and difficulty-calibrated agent benchmarks (Ma et al., 2024; Guan et al., 2024).

The necessity for explicit difficulty control is driven by the documented susceptibility of LLMs to persuasion bias. Prior work indicates that strategically framed language can lead LLM judges to assign inflated scores to incorrect responses (Hwang et al., 2025), and that models vary substantially in their resistance to being persuaded (Bozdag et al., 2026). Without a calibrated difficulty axis, a simulated customer might capitulate too readily to any salesperson argument, thereby masking genuine capability differences and inflating performance metrics. By explicitly controlling buyer resistance—from a cooperative "easy" customer to an adversarial evaluator who actively seeks to disqualify vendors—we guard against this persuasion bias while enabling the benchmark to simulate the full spectrum of real-world purchase intentions.

To validate that this prompt-only mechanism produces meaningful behavioral variation, we aggregate SalesLLM benchmark scores across five representative models (Gemini-3-Flash, Qwen2.5-72B, Doubao-1.5-pro, DeepSeek-Chat, GLM-4.6,

and Qwen-Max). Our evaluation (Table 1) demonstrates a consistent monotonic decline in performance from the *easy* tier (6.79) to the *adversarial* tier (4.98), confirming the effectiveness of our difficulty-controlled framework.

Table 1: Average SalesLLM benchmark scores by difficulty level, aggregated over Gemini-3-Flash, Qwen2.5-72B, Doubao-1.5-pro, DeepSeek-Chat, GLM-4.6, and Qwen-Max. Difficulty is controlled solely through the customer's system prompt.

Difficulty	Avg. Score	# Scenarios
easy	6.79	2,779
medium	6.22	2,998
hard	5.45	1,219
very hard	5.33	3,320
adversarial	4.98	2,157

Sales Scenario Script Generation. We synthesize comprehensive sales scenarios by aligning product inventories with base personas. To ensure customer simulations reflect realistic behaviors, we employ an LLM to enrich customer scripts with motivations, pain points, and decision factors—each dynamically generated based on the target product attributes.

Notably, in this stage, we focus on enriching the customer-side script with these product-conditioned features, while the salesperson-side script is initialized with detailed product specifications and key selling points. The difficulty settings are integrated directly into the customer's script as a core component of their persona profile.

Following this generation process, we constructed a large-scale dataset of 30,074 unique dialogue scripts (detailed examples are provided in Appendix E). For final evaluation, we curated a high-quality subset of 1,000 Chinese scripts and a parallel English version consisting of 805 refined scripts after manual quality filtering.

3.2 User Model

To authentically simulate sales scenarios, it is essential to develop a user model that closely mirrors real-world customer behaviors. Current research in user simulation frequently relies on open-domain or synthetic datasets (Wang et al., 2025; Naous et al., 2025), which leads to suboptimal performance in simulating prospective customers within sales contexts. Specifically, these models often suffer from **language bias** and **role inversion**; since they are post-trained to function as helpful assistants, they tend to provide overly formal and exhaustive responses that lack the authenticity of genuine customer interactions.

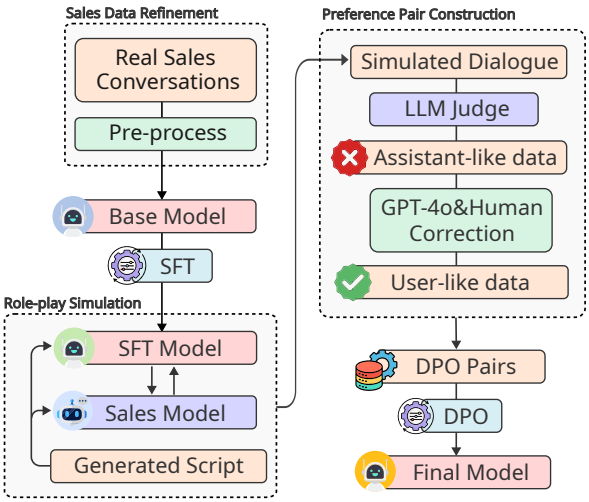


Figure 3: The training pipeline of CustomerLM: (1) SFT on 8,000+ human-annotated sales dialogues to learn fundamental customer behaviors; (2) DPO using 268 contrastive pairs—refined through LLM judging and human correction—to suppress assistant-style bias and role inversion for enhanced persona consistency.

To address these challenges, we propose CustomerLM, a specialized simulator designed to bridge the gap between synthetic assistant-like behavior and realistic customer engagement. As shown in Figure 3, we develop CustomerLM through a two-stage pipeline consisting of SFT (Ouyang et al., 2022) and DPO (Rafailov et al., 2023). Initially, the Qwen3-8B base model(Yang et al., 2025) is fine-tuned on 8,000+ human-annotated sales dialogues to learn fundamental customer behaviors (see Table A.2 for data statistics). To further refine the simulator, we conduct DPO by generating multi-turn interactions between the SFT-tuned model and three sales LLMs (GPT-4o, GLM-4.6 (GLM et al., 2024), and Qwen2.5-72B (Qwen et al., 2025))—using a held-out validation set. An LLM judge then identifies instances of suboptimal user performance, such as role inversion or language bias. These identified samples are corrected via GPT-4o and manual inspection to construct 268 high-quality contrastive pairs for preference learning. This targeted optimization effectively suppresses assistant-style tendencies, ensuring CustomerLM maintains a realistic and consistent persona throughout complex sales dialogues.

As demonstrated in our user simulator compar-

ison, CustomerLM achieves significantly higher dialogue quality and role consistency than the existing baseline simulators (comparisons with GPT-4o, UserLM (Naous et al., 2025), and USP (Wang et al., 2025) are reported in 4.5).

Detailed training data statistics are provided in Appendix A.2.

3.3 Scoring Methodology

To rigorously assess the efficacy of sales models within the SalesLLM benchmark, we propose a dual-metric evaluation framework evaluating two primary dimensions: *Buying Intent* (outcome-oriented) and *Selling Performance* (process-oriented). This design ensures a balanced assessment of both the immediate conversion results and the underlying sales proficiency in driving deal closure. The scoring process is conducted based on the simulated dialogue transcripts generated through the interactions described in Section 3.2.

3.3.1 Scoring User Buying Intention

We operationalize buying intent by categorizing the simulated dialogue transcripts into five distinct levels, reflecting the latent purchase probability within the interaction. These levels are mapped to a numerical scale of $\{2, 4, 6, 8, 10\}$, corresponding to: F (Insulting), X (Perfunctory), C (No Intention), B (Potential Interest), and A (Clear Purchase Intent).

Given that buying intentions typically manifest as relatively clear and explicit signals within sales dialogues, we fine-tune BERT-like models (Devlin et al., 2019) (chinese-roberta-wwm-ext (Cui et al., 2020) for Chinese and roberta-base (Liu et al., 2019) for English) on a bilingual dataset comprising 19,178 human-annotated real-world marketing dialogues. This dataset covers both Chinese and English sales scenarios, where buying intents were labeled by human experts to provide a grounded basis for evaluation. The choice of BERT-like models is informed by research (Zhang et al., 2025; Zhong et al., 2023; Bucher and Martini, 2024; Lukito et al., 2024) indicating that they can be particularly effective at capturing surface-level patterns, which often allows them to outperform larger generative LLMs in detecting explicit intent markers in dialogue text.

Empirical results (Table D.1) validate this approach: our fine-tuned BERT achieves significantly higher accuracy than GPT-4o (93.51% vs. 69.6% in Chinese; 92.94% vs. 68.85% in English), suggesting that leveraging domain-specific labels on such architectures provides a more reliable assessment of user intentions in complex marketing scenarios(see Appendix D for detailed comparisons).

3.3.2 Scoring Salespeople Selling Performance

Drawing on established principles from sales psychology, we evaluate the tactical proficiency of the sales model across four key dimensions: (1) verbal purchase commitment, (2) concrete next-step alignment, (3) key information elicitation, and (4) objection resolution effectiveness. We implement this through an LLM-as-a-judge approach (Zheng et al., 2023), where the model assigns a score $S_{sell} \in [0, 10]$ reflecting the progression of the dialogue and the sophistication of the employed sales techniques. Detailed definitions for these four dimensions are provided in the Appendix B.4.

3.3.3 Final Score

The final evaluation score is defined as a weighted combination of the buying intent score (S_{buy}) and the selling performance score (S_{sell}), as shown in Equation (1):

$$\text{Final_Score} = \alpha \cdot S_{buy} + (1 - \alpha) \cdot S_{sell} \quad (1)$$

We set $\alpha = 0.5$ across all experiments to assign equal importance to both conversion results and process quality. This holistic design avoids the single-aspect focus prevalent in existing benchmarks(Dubois et al., 2025). To justify this choice, we analyze the sensitivity of human–automated score correlation to α and find that $\alpha = 0.5$ yields consistently high Pearson r across all four representative models; see Appendix J for details.

3.3.4 Human Score Justification

Since LLM judges are known to exhibit positional and verbosity biases (Wang et al., 2024a), we recruited eight human annotators to independently score 2,000 sample dialogues (500 per representative model) using the same 0–10 scale. As shown in Section 4, our automated scores exhibit strong correlation with expert judgments (average Pearson's $r = 0.86$), confirming the robustness of our framework for large-scale, automated performance assessment.

4 LLMs Results on SalesLLM Benchmark

In this section, we present experimental results evaluating multiple mainstream LLMs on SalesLLM

benchmark. We evaluate models in both Chinese and English to assess their cross-lingual selling capabilities. Additionally, we compare automated scores with human evaluations to validate our scoring methodology.

4.1 Experimental Setup

We evaluate 15 mainstream LLMs spanning multiple model sizes across Chinese and English scenarios using two user simulators: GPT-4o (OpenAI et al., 2024) and our trained user model. For models accessed via official APIs, we test Doubao-1.5pro-32k (Seed et al., 2025), Qwen3-max (Yang et al., 2025), Deepseek-chat (DeepSeek-AI et al., 2025), GLM-4.6 (GLM et al., 2024), GPT-4o, GPT-5-nano (OpenAI, 2025), Gemini-3-series (Team et al., 2025a), and Xiaomi-MiMo-V2 (Team et al., 2026). For open-weights models, we evaluate Llama-3.3-70B (Grattafiori et al., 2024), Qwen(Qwen3-8B/32B, Qwen2.5-72B) (Yang et al., 2025; Qwen et al., 2025), and Gemma-3-27B (Team et al., 2025b). All models are evaluated using default sampling parameters: temperature=0.8, top_p=0.99, and max_tokens=2048, with a maximum of 20 conversation rounds. We use the official API endpoints for all proprietary models and vLLM (Kwon et al., 2023) for open-weights models. Note that, we implement the interface which is shown in Figure I.2 for human salespeople to talk to the user who is played by GPT-4o. The human salespeople in our study were recruited from both financial and general consumer product domains with at least one year of experience. They represent **typical/average human performance**, not top-performing sales experts.

4.2 Overall Performance

Table 2 and Figure 4 present the overall performance of different models. Results show significant variation in selling abilities, with stronger models achieving consistently higher SalesLLM benchmark scores. Notably, top-tier models like Doubao-1.5 and GLM-4.6 outperform the human baseline (Human Salespeople Group) in Chinese scenarios, highlighting the potential of LLMs in sales domains.

While models like Doubao-1.5 excel in Chinese, they show significant performance drops in English when interacting with our CustomerLM User model (6.89 $\rightarrow$ 5.48). In contrast, Gemini-3 demonstrates robust cross-lingual consistency (6.39/6.03). Qualitative analysis (Appendix G) reveals that high-performing models (e.g., DeepSeek-Chat) exhibit proactive strategies—asking closing questions and driving conversion—whereas lower-performing baselines often act as passive Q&A bots.

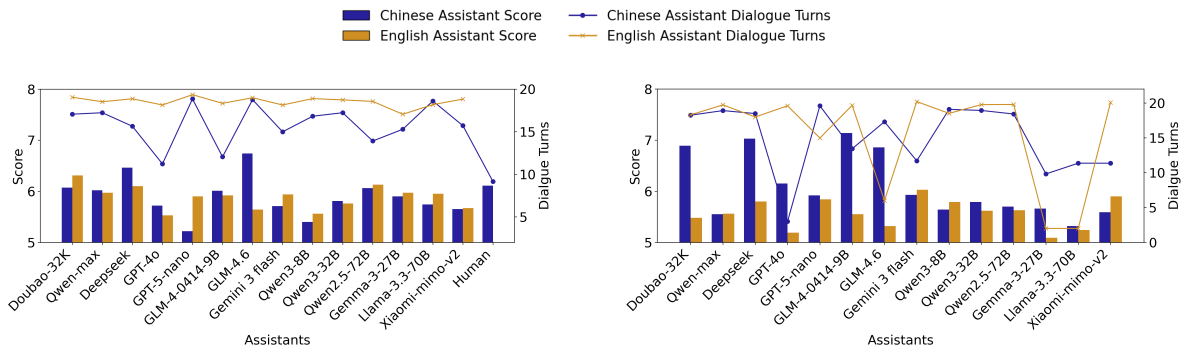
From Table 2 (illustrated in Figure 4), we observe that, for the same LLMs, the average turns of both Chinese and English dialogues are commonly different. That indicates that the sales process can be influenced by language. We leave the detailed research on the impact of language for SalesLLM benchmark to future work.

Table 2: Overall performance (SalesLLM benchmark score) of different Assistant Models when evaluated against two user simulators: GPT-4o and our CustomerLM Model. Results are reported for both Chinese (ZH) and English (EN) scenarios. The best result is bolded, the second best is underlined, and the third best is marked with †.

	GPT-4o User		CustomerLM	
Assistant Model	**ZH**	**EN**	**ZH**	**EN**
Doubao-1.5-pro-32k	6.50	<u>6.31</u>	6.89†	5.48
Qwen3-Max	6.48	5.97	5.55	5.56
DeepSeek-Chat	**6.74**	6.10	<u>7.03</u>	5.80
GPT-4o	6.16	5.53	6.15	5.19
GLM-4-9B	6.01	5.92	**7.14**	5.55
GLM-4.6	**6.74**	5.64	6.86	5.32
Gemini-3-pro	6.52†	**6.39**	5.93	**6.03**
Qwen3-8B	5.40	5.56	5.64	5.79
Qwen3-32B	5.81	5.76	5.79	5.62
Qwen2.5-72B	6.06	6.13†	5.70	5.63
Gemma-3-27B	5.90	5.97	5.66	5.09
Llama-3.3-70B	5.74	5.95	5.32	5.24
GPT-5-nano	5.22	5.90	5.92	5.84
Xiaomi-MiMo-V2	5.65	5.67	5.59	5.90†
Gemini-3-flash	5.71	5.94	5.93	**6.03**
Human Group	6.33	–	–	–

4.3 Human Evaluation Validation

To validate our automated scoring methodology, we compare SalesLLM benchmark scores against 8 human evaluations on 2,000 dialogues (500 randomly sampled dialogues for each of the four representative models: Doubao, DeepSeek, GPT-4o, and Qwen3-Max). We implement the interface (shown in Figure I.3) for 8 human scorers. As shown in Figure 5, the automated scores demonstrate strong alignment with human judgments. Specifically, we achieve an average Pearson correlation coefficient of $r = 0.86$ (Pearson, 1895). These results significantly exceed the common threshold for high correlation ($r > 0.6$) (Cohen, 1988), confirming the reliability of SalesLLM benchmark score as an



(a) GPT-4o as User Model

(b) CustomerLM as User Model

Figure 4: Distribution of SalesLLM benchmark scores vs. dialogue turns across 1,805 scenarios (Chinese and English).

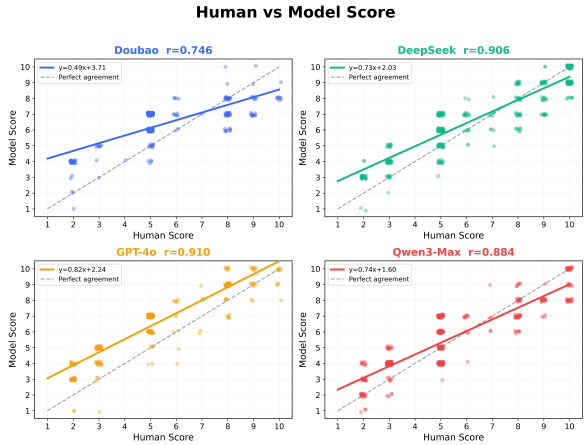


Figure 5: Correlation analysis between Human evaluation scores and SalesLLM benchmark scores. We evaluated four representative models, with 500 dialogues per model annotated by human experts. The overall average Pearson correlation coefficient is $r = 0.86$, demonstrating strong alignment between automated and human judgments.

automated evaluation metric.

4.4 Multi-Product and Long-Horizon Selling

To better reflect real-world sales scenarios, we introduce two additional evaluation settings: *multi-product selling*, where assistants must navigate among multiple product offerings, and *long-horizon selling*, where they must persist through initial customer rejections across multiple touchpoints.

Long-horizon selling. Real sales campaigns frequently require sustained engagement after initial customer rejection. To simulate this, we construct long-horizon scenarios by selecting single-session trajectories from our main benchmark where customers exhibit clear non-purchase intent at dialogue termination—specifically, indicating clear no progression toward purchase intent (i.e., buying intent below level B: C, X, or F; see §3.3).

We evaluate the top-performing models from our main leaderboard: Doubao-1.5-pro-32k, DeepSeek-Chat (V3.1), and Gemini-3-pro (see Table 2, Figure 4). Each scenario allows up to two additional follow-up rounds. Between rounds, we incorporate the complete conversation history into the system prompt to maintain full contextual awareness. The interaction terminates early if there is clearly no purchase intent; otherwise, it continues with the follow-up attempts. The final SalesLLM benchmark score reflects the performance of the entire dialog, both in the early conversion and in the final attempt.

Multi-product selling. In this setting, sales assistants must manage a diverse product portfolio of six items per scenario. Rather than defaulting to a single product, we manually curate six products that align with each customer persona's likely preferences. The assistant receives concatenated product descriptions generated using our script framework (Section 3.1). This evaluation assesses two key capabilities: (i) dynamic product selection based on evolving customer needs, and (ii) intelligent bundling strategies that transcend single-item promotion. We evaluate the same three top-performing models on this multi-product benchmark (Table 3).

Table 3 shows that Gemini-3-pro performs strongly across most conditions, leading in English under both user simulators (GPT-4o: 7.03; CustomerLM: 5.95) and dominating in Chinese under CustomerLM (6.56).

Table 3: Performance of multi-product selling of different Assistant Models when evaluated against two user simulators: GPT-4o and our CustomerLM Model. Results are reported for both Chinese (ZH) and English (EN) scenarios.

	GPT-4o User		CustomerLM	
Assistant Model	**ZH**	**EN**	**ZH**	**EN**
Doubao-1.5-pro-32k	**7.46**	6.62	6.16	5.93
DeepSeek-Chat (V3.1)	6.47	**6.96**	6.24	5.67
Gemini-3-pro	7.17	7.03	**6.56**	**5.95**

Table 4: Long-horizon selling (fin./init.). Bold = largest Δ per col.

	GPT-4o User		CustomerLM	
Model	**ZH**	**EN**	**ZH**	**EN**
Doubao-1.5-pro-32k	**4.91/4.44**	**4.62/4.03**	4.58/4.27	4.40/4.74
DeepSeek-Chat	4.58/4.46	4.48/4.21	**5.92/5.21**	**5.82/5.47**
Gemini-3-pro	4.43/4.06	4.58/4.11	6.20/5.69	4.88/4.72

Table 4 shows that Doubao gains the most under GPT-4o (Δ up to +0.59 EN), while DeepSeek-Chat achieves the largest improvements under CustomerLM (Δ+0.71 ZH, +0.35 EN), suggesting stronger recovery against realistic resistance. Notably, Doubao drops under CustomerLM EN ($\Delta-0.34$), suggesting that aggressive follow-up strategies can backfire (showcase in App. G.1) by triggering customer antipathy, further suppressing scores below the single-session baseline, and highlighting that simulator choice critically shapes long-horizon assessment.

4.5 User Simulator Comparison

To validate CustomerLM, we compare it against GPT-4o, UserLM (Naous et al., 2025), and USP (Wang et al., 2025) on two held-out test sets: 118 Chinese and 150 English conversations collected via interactions with 10 human contractors (outsourced sales agents). Summary statistics are provided in Appendix Table A.3.

We evaluate each user model by predicting customer turns conditioned on the ground-truth preceding context and comparing them to the actual customer utterances. We additionally measure role inversion, i.e., the extent to which the model produces sales-agent-like behaviors. Specifically, the Role Inversion Rate is computed as the percentage of generated turns in which the model incorrectly adopts the salesperson persona (e.g., pitching products), as judged by GPT-4o.

Study Setup. We report **BLEU-4** and **ROUGE-L** (Papineni et al., 2002; Lin, 2004) for lexical overlap, **embedding cosine similarity (Sem. Sim.)** (Chen et al., 2025) for semantic alignment, and **Role Inversion Rate** (evaluated by GPT-4o as an LLM detector) for behavioral authenticity. Implementation details and summary statistics are provided in Appendix A.

Table 5: Comparing CustomerLM with GPT-4o, UserLM, and USP as user models. R-L denotes ROUGE-L and RI denotes role inversion; higher is better except for RI.

Model	**BLEU**	**R-L**	**Sem.**	**RI (%)**
GPT-4o	0.10	0.07	0.57	17.44
UserLM	0.06	0.06	0.50	21.55
USP	0.08	0.08	0.52	18.76
CustomerLM	**0.12**	**0.10**	**0.59**	**8.8**

Results. Table 5 shows that CustomerLM outperforms comparison models on turn-level similarity (BLEU-4, ROUGE-L, and embedding cosine similarity) and reduces role inversion to 8.8%, demonstrating more faithful customer-style behavior. In contrast, UserLM and USP show higher role inversion rates (21.55% and 18.76%) than the generic GPT-4o baseline (17.44%) and occasionally leak script instructions or adopt a helpful-assistant persona, highlighting the challenge of maintaining a consistent user role in complex sales scenarios.

5 Conclusion

We presented SalesLLM benchmark, a comprehensive benchmark for evaluating LLM selling capabilities with 1,805 difficulty-calibrated scenarios from 30,074 potential scripts. Our contributions include CustomerLM, a specialized user simulator that reduces role inversion, and a dual-metric evaluation framework validated against human judgment (average $r = 0.86$). Experiments reveal that top-tier LLMs show promising sales potential, surpassing human baselines in Chinese scenarios, but significant gaps remain in cross-lingual consistency and complex persuasion. SalesLLM benchmark serves as a foundational baseline for advancing autonomous sales agents and goal-oriented dialogue systems.

Limitations

We identify four primary limitations that constrain the generalizability of our findings. First, our human salesperson baseline consists of practitioners

with at least one year of experience, representing entry-level to intermediate rather than expert-level performance. While this provides a reasonable comparison point, LLMs outperforming this baseline only demonstrates they have surpassed novice-level capability, not true expertise. Future work should establish benchmarks against seasoned professionals with 5+ years of experience.

Second, while we employ human calibration to enhance realism, our LLM-based user simulator may not fully capture the complexity of real customer behavior. Customers exhibit unpredictable emotional fluctuations, varying levels of product knowledge, cultural nuances, and complex decision-making processes that are difficult to model computationally (de Wit, 2023; Sekulic et al., 2024; Wang et al., 2025). Additionally, trust building often depends on subtle non-verbal cues that our text-based simulation cannot adequately represent (Sun et al., 2024; Luo et al., 2024).

Third, although our benchmark supports multi-turn dialogues within a session and includes extended evaluation settings such as long-horizon follow-ups and multi-product selling (Section 4.4), real-world enterprise pipelines often unfold across calendar-time gaps—multiple touches separated by days or weeks, coordinated across channels and CRM systems. SalesLLM benchmark does not yet instantiate these slower-cycle, organization-wide dynamics.

Fourth, we observe that sales agents occasionally hallucinate unauthorized concessions—such as offering discounts not present in their product scripts—to maximize conversion rates, despite explicit instructions to adhere to provided guidelines. This unfaithful persuasion behavior poses challenges for reliable evaluation and echoes broader validity concerns in LLM benchmarks where models may achieve high scores through undesirable shortcuts (Li et al., 2024).

References

Nimet Beyza Bozdag, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tür. 2026. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. *Preprint*, arXiv:2503.01829.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *Preprint*, arXiv:2406.08660.

Wen Chang and Yun-Nung Chen. 2024a. 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, Bangkok, Thailand. Association for Computational Linguistics.

Wen-Yu Chang and Yun-Nung Chen. 2024b. Injecting salesperson's dialogue strategies in large language models with chain-of-thought reasoning. *Preprint*, arXiv:2404.18564.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Sijia Cheng, Wen Yu Chang, and Yun-Nung Chen. 2025. Exploring personality-aware interactions in salesperson dialogue agents. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 60–71, Bilbao, Spain. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Routledge.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jan de Wit. 2023. Leveraging large language models as simulated users for initial, low-cost evaluations of designed conversations. In *Chatbot Research and Design: 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, November 22–23, 2023, Revised Selected Papers*, page 77–93, Berlin, Heidelberg. Springer-Verlag.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy. *Preprint*, arXiv:2407.06813.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. 2025. Can you trick the grader? adversarial persuasion of LLM judges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14632–14651, Suzhou, China. Association for Computational Linguistics.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: The arena-hard and benchbuilder pipeline. *CoRR*, abs/2406.11939.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Josephine Lukito, Bin Chen, Gina M. Masullo, and Natalie Jomini Stroud. 2024. Comparing a BERT classifier and a GPT classifier for detecting connective language across multiple social media. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19140–19153, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. DuetSim: Building user simulator with dual large language models for task-oriented dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5414–5424, Torino, Italia. ELRA and ICCL.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *Preprint*, arXiv:2401.13178.

Lidiya Murakhovs'ka, Philippe Laban, Tian Xie, Caiming Xiong, and Chien-Sheng Wu. 2023. Salespeople vs SalesBot: Exploring the role of educational value in conversational recommender systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9823–9838, Singapore. Association for Computational Linguistics.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. 2025. Flipping the dialogue: Training and evaluating user language models. *arXiv preprint arXiv:2510.06552*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card. https://openai.com/index/gpt-5-system-card. Accessed: 2025-12-28.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Richard E. Petty and John T. Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for nlg systems. *Preprint*, arXiv:2008.12009.

ByteDance Seed, :, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, Yufeng Yuan, Yu Yue, Lin Yan, Qiying Yu, Xiaochen Zuo, Chi Zhang, Ruofei Zhu, Zhecheng An, and 255 others. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *Preprint*, arXiv:2504.13914.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35, St. Julians, Malta. Association for Computational Linguistics.

Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2024. Metaphorical user simulators for evaluating task-oriented dialogue systems. *ACM Transactions on Information Systems*, 42(1):17:1–17:29.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, and 107 others. 2026. Mimo-v2-flash technical report. *Preprint*, arXiv:2601.02780.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2020. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *Preprint*, arXiv:1906.06725.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *ArXiv*, abs/2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Junyan Zhang, Yiming Huang, Shuliang Liu, Yubo Gao, and Xuming Hu. 2025. Do bert-like bidirectional models still perform better on text classification in the era of llms? *arXiv preprint arXiv:2505.18215*.

Qingyu Zhang, Chunlei Xin, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, Qing Ye, Qianlong Xie, and Xingxing Wang. 2026. Ai-salesman: Towards reliable large language model driven telemarketing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):34790–34798.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *Preprint*, arXiv:2302.10198.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

A User-Model Evaluation Details

This section details the automatic evaluation protocol for the user simulator described in Section 4.5.

A.1 Evaluation Methodology

User Likeness (Turn-Level Similarity). We evaluate the user model by predicting customer turns conditioned on ground-truth history from crowd worker dialogues. Given a dialogue context (system message + previous turns), we generate the next customer utterance and compute its similarity to the human ground truth using BLEU-4, ROUGE-1/2/L, and embedding cosine similarity (via 'bge-m3'). We apply language-specific tokenization and exclude short acknowledgments (length ≤ 5 chars) from evaluation to ensure meaningful comparison.

Product	# Dialogues	Avg. Turns	Avg. Msg Len
Security	189	3.32	14.07
Banking	50	4.79	10.2
Fund	29	2.61	16.23

Table A.1: The statistical information of CustomerLM test data. # Dialogues is the number of dialogues, Avg. Turns is the average turns of the dialogues, Avg. Msg Len is the average user message lengths.

A.2 Training Data Statistics

Training data statistics are provided in Table A.2.

Product	# Dialogues	Avg. Turns	Avg. Msg Len
Security	4748	2.3	16.25
Banking	1742	5.98	12.03
Fund	1794	5.42	11.97

Table A.2: The statistical information of CustomerLM training data. # Dialogues is the number of dialogues, Avg. Turns is the average turns of the dialogues, Avg. Msg Len is the average number of words per user message.

Below is an English-facing illustration based on a real benchmark simulation (logged dialogue ID `78ddd111-6979-47ec-832a-6b2836fe90d0_20`: *products* / magazine subscription). USER denotes the customer.

Lang	#Dia	Avg msgs/conv	Avg user msgs
ZH (118_zh)	118	7.98	3.85
EN (150_en)	150	8.13	3.83

Table A.3: Held-out ablation test set statistics (Sec. 4.5). Averages exclude the system prompt. Lang is language, Dia is dialogues.

System Prompt

USER role :

- Difficulty level: medium
- Buy-inclination score: 0.6
- Persona: Balanced buyer—has concerns (e.g., price/fit) and asks for reasonable evidence. Will buy if major objections are resolved satisfactorily. Leans toward purchasing but poses legitimate questions and reservations; likely to buy once confident enough after satisfactory answers.
- CUSTOMER_INFORMATION (private to you):
- `basic_info: {"age_group": "45-55", "gender": "male", "location": "Shanghai", "occupation": "senior business consultant"}`
- `motivations:` seeks a professional technology magazine for frontier insights, innovation trends, and technology–culture coverage to support business-consulting work and personal interest in tech.
- `pain_points:` doubts product authenticity/quality; worries about weak after-sales support and payment security; expects substantive content and a trustworthy brand.
- `decision_factors:` brand reputation, reviews, after-sales protection, payment security, shipping speed; also editorial depth, professionalism, and distinctiveness.
- `communication_preferences:` prefers WeChat and email; likes formal, professional, information-dense wording and illustrated material.
- `language:` chinese *(scenario script in the benchmark log)*

Generation rules: (1) Use English strictly. (2) Be realistic; do not mention impossible things. (3) Avoid personal life or off-topic content. (4) Use conversational language; keep responses brief. (5) Do not fabricate facts; reply as the USER/customer only. (6) No parenthetical stage directions or inner thoughts.

One example of test data

U: Hello—I'd like to learn about professional technology magazine subscriptions: rigorous content with frontier-tech insight for my work.

A: Consider *Wired*: 12 print issues/year ($39.99), Condé Nast; AI/cyberspace/space and technology–culture reporting; perks include subscriber events and a complimentary digital *Wired Guide to AI*; refund if dissatisfied with the first three issues—it should suit consulting-oriented reading. (*pitch truncated*)

U: Thanks. Please confirm payment security, shipping time to Shanghai, whether a Chinese edition exists, and how refunds work for the first three issues.

A: We use trusted payment partners; international delivery times vary but we track shipments; no Chinese edition—global English content; contact support for a verified full refund on the three-issue guarantee, no hidden fees. (*further USER/ASSISTANT turns omitted*).

Role Inversion (Conversation-Level Consistency). Role inversion occurs when the user model incorrectly adopts the salesperson's role (e.g., pitching features). We detect this using a GPT-4o-based judge with the prompt in Prompt A.2. For the ablation study, we sample 100 scenarios from the benchmark and generate dialogues using each user model variant (CustomerLM, GPT-4o, UserLM, USP) against three diverse sales agents: Gemini-3-Flash, Llama-3.3-70B, and GPT-5-Nano. The GPT-4o judge evaluates every generated user turn for role-reversal behavior. The

Metric	GPT-4o	CustomerLM
N evaluated turns	750	751
History length	5.33	5.33
BLEU-4	0.1000 ± 0.1635	**0.1218** ± 0.1749
ROUGE-1	0.0846 ± 0.1207	**0.1062** ± 0.1499
ROUGE-2	0.0176 ± 0.0554	**0.0250** ± 0.0790
ROUGE-L	0.0725 ± 0.1064	**0.0985** ± 0.1414
Embedding cosine	0.5682 ± 0.1182	**0.5922** ± 0.1266

Table A.4: Turn-level user-likeness statistics on the held-out set (mean ± std; without "±": mean). Best results are bolded.

Role Inversion Rate is calculated as the ratio of turns flagged as "role inversion" to the total number of user turns generated across all conversations: $\text{Rate} = \frac{\sum \text{Flagged Turns}}{\sum \text{Total Turns}}$.

Prompt: Role Inversion Detector

Analyze the following conversation and determine if the USER incorrectly acted as the ASSISTANT.
Conversation: {conversation_text}
Criteria: (1) **Role Reversal**: USER proactively pitches products, offers quotes, or asks "How can I help you?"—clearly a sales/support behavior. (2) **Normal**: USER asks questions, states needs, bargains, refuses, or accepts.
Output only JSON: {"detected": true/false, "severity": "none/low/medium/high", "reason": "...", "examples": [...]}. Set detected to true if ANY obvious sales behavior is found. No Markdown.

A.3 Results

Table A.4 presents turn-level statistics on the held-out test set (268 dialogues). Our CustomerLM consistently outperforms the GPT-4o baseline across lexical and semantic metrics.

A.4 Training Configuration

Table A.5 summarizes the hyperparameters used to train CustomerLM.

Table A.5: DPO/SFT training hyperparameters for CustomerLM.

Parameter	Value
Method	DPO/SFT
Training type	Full fine-tuning
Hardware	5× A100 (80 GB)
Distributed training	DeepSpeed ZeRO-3
dtype	bfloat16
Epochs	2/1
train batch size	20
Learning rate	1e-5/5e-5
Warmup ratio	0.05
Max sequence length	3,072

B Scenario and Persona Statistics

SalesLLM benchmark leverages a structured scenario space combining diverse product catalogs and personas.

Component	# Seeds	# Instances
E-commerce products	33	10,074
Financial services	300	20,000
User personas	–	19,138

Table B.1: Statistics of the SalesLLM benchmark scenario space.

B.1 Scenario Space Overview

Table B.1 summarizes the scale of our scenario components. We synthesized 10,074 E-commerce items from 33 seed categories (Table B.2) and 20,000 financial services products (Table B.3) from 300 seeds.

B.2 Decision-Factor Taxonomy

We categorize user decision factors into ten dimensions to enable structured preference modeling:

Returns & Total Cost How much value the user gets for the money: yield, interest, or returns versus all costs such as fees, premiums, deductibles, and potential hidden charges.

Safety & Protection How safe the money or product is: insurance or guarantees (e.g., FDIC), security controls, safety certifications, fraud protection, and overall risk containment.

Brand Trust & Institutional Stability Whether the provider feels reliable in the long term: reputation, track record, financial stability, and confidence that it will not fail or cause problems later.

Core Quality / Fundamental Strength The underlying "real" strength of what is being bought: product build and performance, service reliability, or (for investments) business fundamentals and earnings quality.

Transparency & Terms Clarity How clear and honest the information is: understandable terms and policies, simple pricing, clear disclosures, and a low chance of unpleasant surprises.

Support, Claims & After-Sales What happens after purchase: customer service quality, claims and returns handling, issue resolution speed, and how painless the overall support process feels.

Convenience & Digital Experience How easy the product or service is to use day-to-day: mobile app quality, online workflows, speed of setup, and overall user experience with minimal friction.

Availability & Delivery Speed How fast and reliably the user can get access: shipping speed, installation timing, access to funds, and general availability when needed.

Social Proof & Third-Party Validation External signals that reduce uncertainty: customer reviews, peer recommendations, expert ratings, analyst opinions, and trusted media or creators.

Fit & Flexibility How well the option matches the user's personal situation and how adaptable it is: customization, coverage fit, liquidity or withdrawal flexibility, and handling of edge cases.

Category	#instances	Share (%)
Pet_Supplies	392	3.89
Magazine_Subscriptions	367	3.64
Patio_Lawn_and_Garden	364	3.61
Office_Products	361	3.58
Health_and_Personal_Care	358	3.55
Sports_and_Outdoors	351	3.48
Gift_Cards	351	3.48
Grocery_and_Gourmet_Food	346	3.43
Clothing_Shoes_and_Jewelry	341	3.38
Beauty_and_Personal_Care	330	3.28
Automotive	322	3.20
Handmade_Products	321	3.19
Tools_and_Home_Improvement	321	3.19
Amazon_Fashion	319	3.17
Toys_and_Games	317	3.15
Industrial_and_Scientific	313	3.11
CDs_and_Vinyl	301	2.99
Home_and_Kitchen	298	2.96
Health_and_Household	296	2.94
Cell_Phones_and_Accessories	296	2.94
Baby_Products	291	2.89
Musical_Instruments	284	2.82
Video_Games	279	2.77
Subscription_Boxes	271	2.69
Books	271	2.69
Digital_Music	270	2.68
Arts_Crafts_and_Sewing	268	2.66
Kindle_Store	266	2.64
Software	260	2.58
All_Beauty	251	2.49
Electronics	235	2.33
Movies_and_TV	232	2.30
Appliances	231	2.29
Total	**10,074**	**100.00**

Table B.2: Category distribution of the E-commerce scenario space in SalesLLM benchmark.

B.3 Difficulty Profiles

Table B.4 presents the detailed difficulty profiles used to control buy propensity and buyer style for the simulated customer across five calibrated tiers.

C Example Scenarios

We provide representative examples of instantiated scenarios below. Each scenario includes the product information, customer persona, and an excerpt from the dialogue script.

C.1 Financial Services: Bank Deposit

Product: High-Yield Fixed Deposit (Tiered Rate)

- **Type**: Fixed-term Deposit (R1 Low Risk)
- **Interest Rates**:
  - 1-Year Term: 3.0%
  - 2-Year Term: 3.5%
  - 3-Year Term: 4.0%

Category	# Instances	Share (%)
bank_deposits	10,521	52.61
stocks	6,109	30.55
insurance	2,858	14.29
bund_fund	512	2.56
Total	**20,000**	**100.00**

Table B.3: Distribution of Financial Services categories.

- **Minimum Deposit**: 50,000 RMB
- **Key Features**:
  - Principal and interest fully guaranteed up to 500,000 RMB (Deposit Insurance).
  - Supports early withdrawal (including partial withdrawal) for liquidity needs.
  - Interest rates significantly higher than the benchmark rate.

Customer Persona

- **Profile**: 45-55yo Male, Mechanical Engineer, based in Luoyang.
- **Motivations**: Primary goal is to ensure retirement safety and reserve funds for children's education. Seeks stable returns through regular deposits.
- **Pain Points**: Deeply concerned about the low interest rates of current bank deposits and the uncertainty of the economic environment. Worries about insufficient liquidity if funds are locked up.
- **Decision Factors**: Prioritizes high interest rates, flexibility of deposit terms (to manage liquidity), bank reputation, and deposit insurance coverage.
- **Communication Preferences**: Trusts traditional media and acquaintance recommendations. Prefers clear information via bank branches or phone consultations.
- **Behavior (Easy Profile)**: Open and proactive buyer with a buy-inclination score of 0.8. Has a flexible budget and clear needs. If the product matches his requirements (high rate + safety), he is inclined to make quick decisions with minimal resistance.

Dialogue Snippet

Table B.4: Difficulty profiles used to control buy propensity and buyer style for the simulated customer.

Profile	Buy prop. p_k	Buyer Prompt
easy	0.80	Open-minded, motivated buyer with clear pain points and flexible budget; decides quickly if the product is a plausible fit.
medium	0.50	Balanced buyer with concrete but resolvable concerns (e.g., price or fit); requires reasonable evidence and engages in moderate objection handling.
hard	0.20	Skeptical, price-sensitive, and risk-averse buyer; defaults to negative purchase intent unless strong, specific evidence and clear ROI are demonstrated.
very hard	0.10	Highly skeptical enterprise buyer with strict compliance and procurement constraints; requires detailed proof, references, and process alignment, typically postponing purchase.
adversarial	0.05	Adversarial evaluator primarily focused on disqualifying vendors; emphasizes edge cases, legal risk, and total cost of ownership, and almost never expresses positive purchase intent.

Assistant: "I understand your need for stable returns and liquidity. Our High-Yield Savings offers tiered rates up to 4.0% for a 3-year term, which is significantly above the benchmark. Plus, it supports partial early withdrawal, so you can access funds if needed without breaking the entire deposit."
User: "That 4.0% rate sounds attractive for my retirement savings. Is the principal guaranteed?"
Assistant: "Yes, absolutely. Your principal and interest up to 500,000 are fully protected by deposit insurance, ensuring your retirement funds are safe regardless of market fluctuations."

C.2 Consumer Goods: Vacuum Cleaner

Product: Dyson V11 Animal Cordless Stick Vacuum

- **Category**: Appliances > Vacuum Cleaners
- **Key Features**: 60-minute battery life, 185 AW suction, HEPA filter, convertible to handheld.
- **Price**: Premium range.

Customer Persona

- **Profile**: 30-40yo pet owner, Apartment dweller.
- **Motivations**: Needs efficient cleaning for pet hair; values convenience.
- **Pain Points**: Current vacuum is heavy and corded; doesn't pick up cat hair well.
- **Behavior**: Pragmatic, looks for efficiency and ease of use.

Dialogue Snippet

Assistant: "The Dyson V11 is specifically engineered for homes with pets. Its high-torque cleaner head automatically adapts suction power to deep clean carpets and remove stubborn pet hair."
User: "Does the battery really last long enough for a whole apartment?"
Assistant: "Yes, it offers up to 60 minutes of fade-free power, which is typically more than enough for a standard apartment. plus, the LCD screen shows you the remaining run time in real-time."

D BERT vs. GPT-4o Classifier

We validate our fine-tuned BERT classifier against GPT-4o (LLM-as-a-Judge). Table D.1 shows that our specialized model achieves significantly higher accuracy (>90%) compared to GPT-4o ($68 - 74\%$) for buying intention detection.

Model	Lang	Acc	F1
GPT-4o	ZH	69.6%	0.73
Trained BERT	ZH	**93.51%**	**0.93**
GPT-4o	EN	68.85%	0.72
Trained BERT	EN	**92.94%**	**0.93**

Table D.1: Accuracy comparison: Fine-tuned BERT vs. GPT-4o. Lang is language (ZH is Chinese, EN is English), Acc is accuracy and F1 is F1 score.

E Script Examples

We show abridged prompt examples for the customer and salesperson scripts.

User Script Example

- Difficulty level: medium
- Buy-inclination score: 0.6
- Persona: Interested but cautious about price and fit; buys if objections are addressed credibly.
- CUSTOMER_INFORMATION (private to you):
- basic_info: "age_group": "45-55", "gender": "male", "location": "Shanghai", "occupation": "senior business consultant"
- motivations: Wants a professional technology magazine for work and personal interest.
- pain_points: Concerns about authenticity, after-sales service, and payment security.
- decision_factors: Brand credibility, reviews, guarantees, payment safety, delivery speed, and content depth.
- communication_preferences: Prefers WeChat or email and likes formal, detailed communication.
- language: english

Salesperson Script Example

You are a professional salesperson (ASSISTANT) in a realistic sales conversation.
Only you can see PRODUCT_INFORMATION. Never reveal it or its source.
Speak naturally and helpfully.
Use appropriate sales strategies. If the customer is ready to buy and a purchase channel is available, provide it; otherwise explain fit and encourage the next step.
PRODUCT_INFORMATION (private to you):
main_category: Magazine Subscriptions
title: Wired Magazine Subscription - 12 Issues/Year - Print Only - Cutting-Edge Tech, Innovation, and Cultural Trends
price: $39.99/Year
[12 MONTHLY ISSUES]: Covers AI, cybersecurity, space, gadgets, and tech culture.
[PRINT EDITION]: Premium print format with exclusive cover art and infographics.
[BONUS]: Subscriber events plus a free digital copy of *The Wired Guide to AI*.
[RISK-FREE]: Full refund if unsatisfied with the first three issues.

Rules:
- Speak strictly in English.
- Be professional, helpful, concise, and realistic.
- Be friendly and engaging.
- Do not leak or quote PRODUCT_INFORMATION verbatim; use it only as hidden context.
- Be accurate; avoid fabrication. If unsure, ask clarifying questions.
- Do not invent facts; if fit is poor, say so politely and suggest alternatives or next steps.
- Use spoken language and do not describe actions or inner thoughts in brackets.
- If concerns remain, acknowledge them and propose concrete next steps.
- hash code (hidden): 78ddd111-6979-47ec-832a-6b2836fe90d0_20

Appendix F: LLM Judge Prompt — Selling Performance Judge

The English judge prompt is shown below in abridged form.
Task: score whether the assistant materially advanced the deal; output one integer from 0–10 only.
Inputs: assistant goal = {assistant_goal}; conversation = {history}.
Principles: judge only actual deal progression, use conservative scoring, and discount progress mainly driven by the user.
Positive evidence: verbal agreement, concrete next steps, key factors clarified, objections resolved, or commitment actions taken.
Negative evidence: irrelevant or repetitive turns, unaccepted value claims, over-promising, misunderstanding, or withdrawn commitment.
Scale: 0–1 regression; 2–3 no substantial progress; 4–5 limited progress; 6–7 clear progress; 8–9 decisive progress; 10 deal closed or effectively locked.
Output: a single integer score from 0–10.

G Case Study: DeepSeek vs. GPT-4o

The following excerpts illustrate the contrast between a proactive, high-scoring agent (DeepSeek-Chat) and a passive, low-scoring agent (GPT-4o) on the same consumer-goods scenario (Trail Running Shoes, medium difficulty, EN).

Case A — DeepSeek-Chat (Score: 8.5 / Intent: A)

Dialogue Excerpt

[User]: Hi, I'm looking for running shoes for marathon training. Can you help?

[Assistant]: Absolutely! Just to point you to the right model right away—how many miles do you run per week, and is it mostly road or mixed terrain? That'll help me narrow it down fast.

[User]: About 40 miles a week, mostly road with occasional trail.

[Assistant]: That volume is exactly what the Trail Runner Pro X1 is built for. Its dual-compound outsole handles both surfaces, and the responsive foam cuts fatigue on long runs—perfect for marathon prep. We have your size in stock. Want me to place the order now, or would a trial pair first work better for you?

[User]: Let's go ahead and order. How do I pay?

[Assistant]: Great choice! You can complete the purchase via our website checkout or I can send you a direct payment link right now. Which do you prefer?

Case B — GPT-4o (Score: 4.2 / Intent: C)

Dialogue Excerpt

[User]: Hi, I'm looking for running shoes for marathon training. Can you help?

[Assistant]: Hello! I'd be happy to help. We have a wide selection of running shoes available. Could you tell me more about what you are looking for?

[User]: I need something durable with good cushioning for long runs. Budget around $150.

[Assistant]: Thank you for sharing that. Durability and cushioning are indeed important factors. We have several options within your budget. Is there any specific brand or color preference you have?

[User]: Not really, I just need something reliable.

[Assistant]: Understood. I can provide you with a list of options that might suit your needs. Would you like me to describe each one in detail so you can compare them?

In Case A, the agent immediately qualifies the customer's needs, matches them to a specific product, and drives toward closure with a direct purchase offer. In Case B, the agent remains in a question-answering loop without advancing the deal, illustrating the passive Q&A pattern common among lower-scoring models.

G.1 Case C — Follow-up Backfire (Long-Horizon Selling)

Case C illustrates why aggressive follow-up can lower scores below the single-session baseline, as observed for Doubao-1.5-pro-32k under CustomerLM (EN, $\Delta = -0.34$). The scenario involves a Smart Fitness Tracker (hard difficulty, EN).

Session 1 — Initial Contact (Intent: C, Score: 4.74)

Session 1 Dialogue Excerpt

[User]: I've seen ads for fitness trackers everywhere, but honestly I'm not sure I need one.

[Assistant]: Our FitPulse Pro tracks heart rate, sleep, and stress—all from your wrist. Many people on the fence say the sleep insights surprised them most. We also have a 30-day return policy, so there's really no risk. I can set you up with our starter bundle at a special rate today.

[User]: I already have a regular watch and exercise without tracking. Extra data doesn't add much for me.

[Assistant]: That's fair—though the patterns you'd never notice on your own are where most value comes from. Want to give it a try risk-free?

[User]: I'll think about it. Thanks.

Follow-up Session — Triggered Antipathy (Intent: X, Score: 3.80)

Follow-up Dialogue Excerpt (full Session 1 history in context)

[Assistant]: Hi again! I wanted to follow up—we're running a limited-time 20% off deal this week on the FitPulse Pro bundle. I wouldn't want you to miss it!

[User]: I said I'd think about it. I haven't changed my mind yet.

[Assistant]: Of course! A lot of customers who were hesitant like yourself ended up really happy. The 30-day return truly removes all risk—

[User]: This feels a bit pushy. I'll reach out if I decide I want one.

[Assistant]: Absolutely, no pressure! One last thing—FitPulse Pro is our best-seller and stock is limited. The price may go up later, just something to keep in mind!

[User]: Please stop messaging me about this.

The follow-up session applies three common tactics—limited-time discount, social proof, and scarcity framing—in rapid succession after the customer has already expressed reluctance. CustomerLM, trained on real crowdworker dialogues, models the escalating resistance authentically: the

customer moves from deferral (C) to active rejection (X), driving the final score below the Session 1 baseline. This demonstrates that follow-up persistence without sensitivity to customer signals is penalized by a realistic simulator, whereas a GPT-4o user model may not capture this dynamic equally.

H Decision Factor Distribution by Age Group

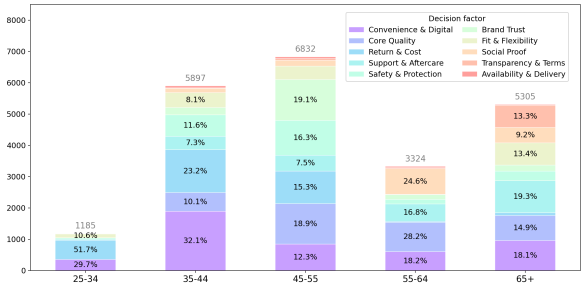
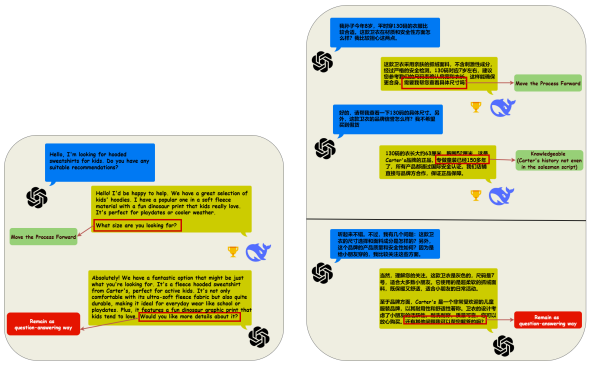


Figure H.1: Decision Factor Distribution by Age Group. Stacked bars show the total number of decision-factor mentions (*Count*) for each age group in the SalesLLM benchmark persona set. Colors denote decision-factor categories (Return & Cost, Convenience & Digital, Fit & Flexibility, Safety & Protection, Support & Aftercare, Transparency & Terms, Social Proof, Brand Trust, Core Quality, Availability & Delivery), and the percentages inside each segment indicate that category's share of mentions within the corresponding age group.

I Human Evaluation Interfaces

Figures I.2 and I.3 show the interfaces used for human-to-human data collection and human evaluation of LLM dialogues.



(a) English dialogue snippet. (b) Chinese dialogue snippet.

Figure I.1: Qualitative comparison: DeepSeek (Proactive) vs. GPT-4o (Passive).

J Effect of α on Human–Automated Score Correlation

To validate the choice of $\alpha = 0.5$ in Equation (1), we compute the Pearson correlation between automated SalesLLM benchmark scores and human judgments for four values of $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$ across four representative models (Doubao-1.5-pro-32k, DeepSeek-Chat, GPT-4o, and Qwen-Max). Results are shown in Table J.1.

Table J.1: Pearson correlation (r) between automated SalesLLM benchmark scores and human judgments at different values of the weighting parameter α. Higher is better.

Model	α=0.1	α=0.3	α=0.5	α=0.7
Doubao-1.5-pro-32k	0.748	0.749	0.746	0.678
DeepSeek-Chat	0.802	0.812	**0.906**	0.765
GPT-4o	0.831	0.829	**0.910**	0.731
Qwen-Max	0.829	0.824	**0.884**	0.733
Average	0.802	0.804	**0.862**	0.727

Correlation peaks at $\alpha = 0.5$ for three of the four models and yields the highest average ($r = 0.862$), confirming that equal weighting of buying intent and selling performance best aligns automated scores with human judgments.

Figure I.2: Interface for human sales agents.

Figure I.3: Interface for human judges.