

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, 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-based rater 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 16.84% (GPT-4o) to 4.33%. SalesLLM scores correlate strongly with expert human ratings (Pearson $r = 0.98$). 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 serves as a scalable benchmark for developing and evaluating outcome-oriented sales agents.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in real-world, goal-directed interactions, with sales emerging as a high-impact application (Hanschmann et al., 2023; Murakhovs’ka et al., 2023; Chang and Chen, 2024; Zhang et al., 2025b; Cheng et al., 2025). Yet most benchmarks emphasize dialogue quality (Liu et al., 2016; Sai et al., 2020) and provide limited coverage of outcome-oriented abilities such as proactive influence and goal achievement. Goal-oriented dialogues are widely studied (Chen et al., 2017; Sun et al., 2021),

but persuasion in realistic business settings remains underexplored.

Sales in the real world represents a distinct form of **asymmetric persuasion**, where an agent seeks to influence user attitudes and drive conversion (Petty and Cacioppo, 1986; Wang et al., 2020) despite potential resistance arising from personal constraints. This goal-oriented dynamic fundamentally differs from general social interaction benchmarks (Zhou et al., 2024; Chen et al., 2024). To address this gap, we introduce **SalesLLM**, a benchmark dedicated to evaluating proactive persuasion and adaptive planning for sales. Leveraging Qwen3-72B (Yang et al., 2025), we synthesize 30,074 task scripts and 19,138 role profiles, curating 1,805 diverse multi-turn scenarios from realistic applications across Financial Services and Consumer Goods, with concrete scenario examples provided in Appendix C. As illustrated in Figure 1, SalesLLM employs a dual-assessment framework: LLM judges evaluate the efficiency of the sales process, while BERT-based models estimate purchase intention as a proxy for persuasive effectiveness.

Realistic user simulation is challenging: general-purpose LLM simulators often exhibit formal language bias and role confusion (Wang et al., 2025; Sun et al., 2024; Luo et al., 2024; Schatzmann et al., 2006; Algherairy and Ahmed, 2025; Sekulic et al., 2024; de Wit, 2023; Chen et al., 2026). To archive realistic selling simulation, we build a specialized user simulator **CustomerLM**, from 8,000 crowdworker-involved dialogues, using Qwen3-14B (Yang et al., 2025) as the agent, and train it with SFT (Brown et al., 2020) and DPO (Rafailov et al., 2023). We then curate 300 high-quality dialogues for robust evaluation.

Our contributions are as follows:

- We propose SalesLLM, a benchmark for sales persuasion with 1,805 multi-turn realistic scenarios in Chinese (1,000) and English (805).

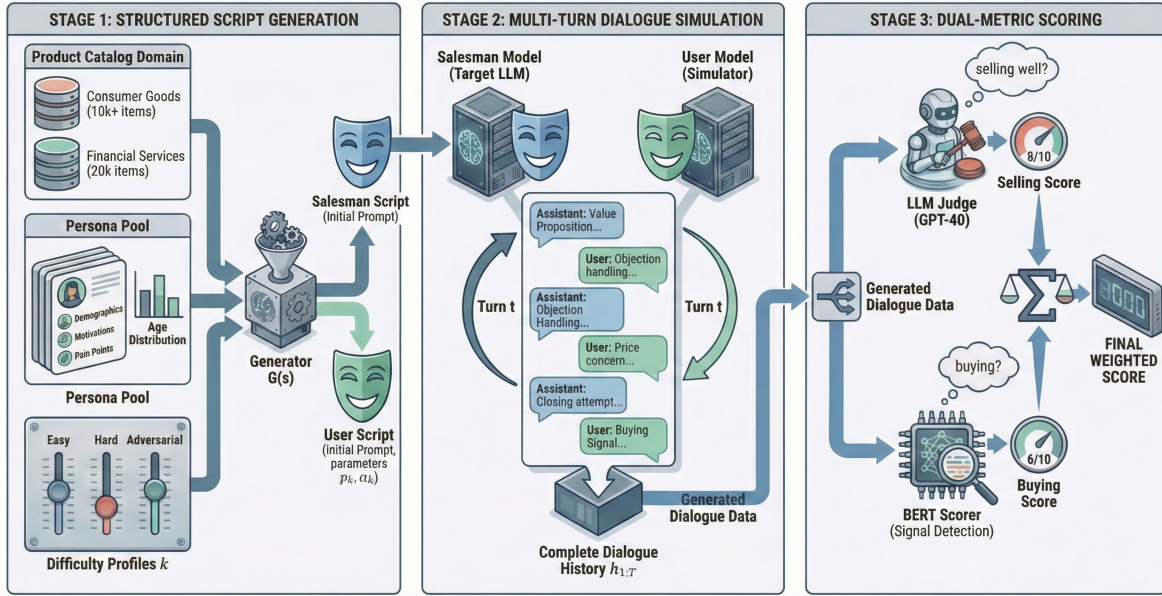


Figure 1: The SalesLLM benchmark pipeline consists of three stages: 1) Script generation: Creating standardized role-play scripts for various contexts; 2) Dialogue simulation: Generating multi-turn dialogues between the target LLM (salesperson) and a virtual customer (GPT-4o or CustomerLM); 3) Scoring: Measuring success based on customer purchase intent and salesperson performance.

- We develop a realistic user simulator, CustomerLM, trained on 8,000 human-involved sales and marketing dialogues using SFT and DPO.
- We design an automated dual-scoring framework combining process efficiency and purchase intention for sales outcome-oriented evaluation.
- Our evaluation results of mainstream LLMs and human participants reveal gaps in high-stakes asymmetric persuasion.

2 Related Work

Multi-turn benchmarks evaluate LLMs’ strategic interaction abilities, including role-play frameworks for persona maintenance and character-driven conversations (Wang et al., 2024; Ma et al., 2024; Liu et al., 2024), as well as benchmarks like Sotopia (Zhou et al., 2024) and DailyPersuasion (Jin et al., 2024) that focus on multi-turn social or persuasive dialogues. While these benchmarks provide insights into general social intelligence and multi-turn dialogue capabilities, they do not explicitly model interactions with asymmetric incentives, explicit conversion objectives, and measurable behavioral outcomes, which are central to sales dialogues. Furthermore, although there are existing efforts in

training LLMs for sales (Hanschmann et al., 2023; Murakhovska et al., 2023; Chang and Chen, 2024; Zhang et al., 2025b; Cheng et al., 2025), they often rely on small-scale evaluation datasets and lack a systematic, large-scale benchmark for assessing sales capabilities.

Existing LLM-based user simulators often rely on general-purpose models to mimic user behavior (de Wit, 2023; Sekulic et al., 2024; Algherairy and Ahmed, 2025). This can produce overly formal or unnatural responses, role confusion, and unrealistic decision-making, limiting the fidelity of evaluation in outcome-driven interactions (Wang et al., 2025). These gaps motivate SalesLLM, a comprehensive and systematic evaluation benchmark for realistic sales scenarios, to introduce a verifiable assessment framework for LLMs in asymmetric persuasion contexts.

3 SalesLLM

In this section, we introduce SalesLLM, a comprehensive benchmark designed to evaluate the strategic selling intelligence of Large Language Models (LLMs) in realistic business scenarios. The whole pipeline of SalesLLM benchmark consists of three stages and is shown in Figure 1. At stage 1, we first construct 30,074 standardized role-play scripts (Section 3.1) across diverse sales contexts. The

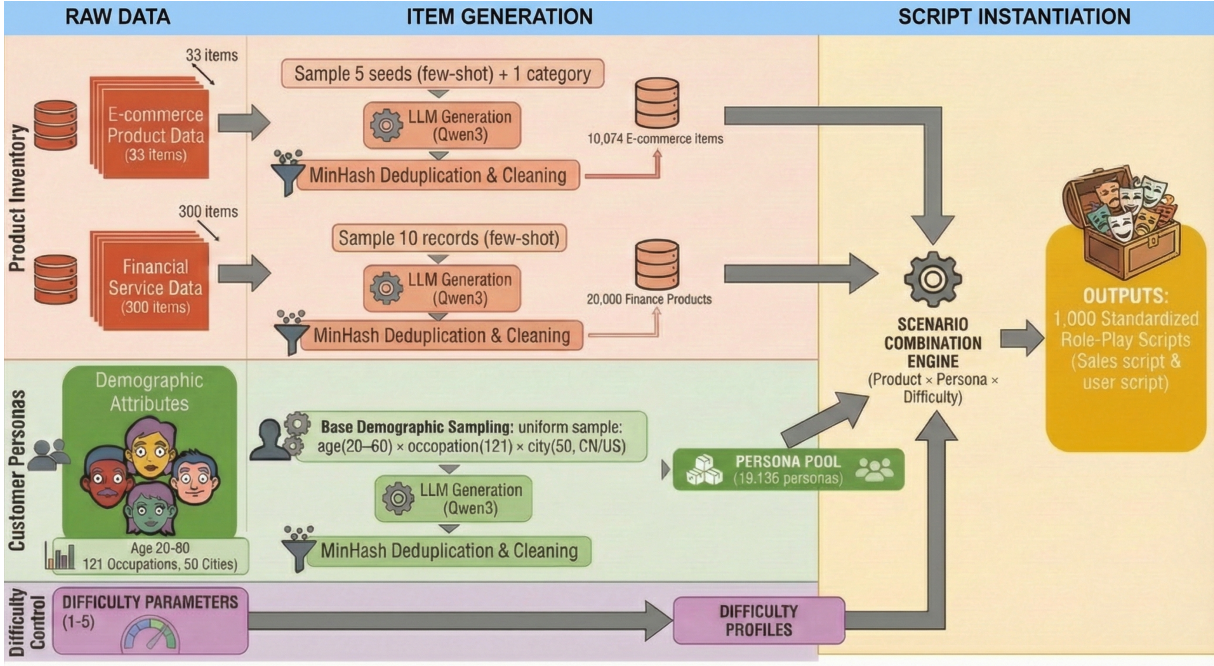


Figure 2: Overview of the SalesLLM Script Generation Pipeline. The pipeline follows a hierarchical process: (i) **Raw Data Stage**: collecting seed items for financial services and consumer goods (Amazon-Reviews-2023) and defining base persona attributes (age, occupation, city). (ii) **Item Generation Stage**: synthesizing detailed product profiles with selling points via few-shot learning and constructing a persona pool. (iii) **Script Instantiation Stage**: performing conditional sampling to enrich personas with product-aligned motivations and pain points, integrated with difficulty controls to produce the final standardized role-play scripts for both the salesperson ($\text{script}^{\text{sales}}$) and the user simulator ($\text{script}^{\text{user}}$).

scripts are the system prompts that define role information for both user and salesperson, and are assigned to the LLMs before a simulation starts. The Appendix F shows an example for customer and salesperson scripts. During evaluation, at stage 2, the target LLM adopts the salesperson role, while the customer role is fulfilled by either GPT-4o (OpenAI et al., 2024) or our specialized user model CustomerLM (Section 3.2). At stage 3, the generated dialogues are evaluated for sales effectiveness, measured by customer purchase intent and salesperson performance (Section 3.3).

3.1 Script Generation

We constructed realistic scripts for role-play-based evaluations utilizing a structured persona modeling approach (Wang et al., 2024; Ma et al., 2024) to define customer (user) profiles, while initializing salesperson (assistant) roles with target product information and key selling points.

Structured scenario space. To enable scalable and diverse script generation, as illustrated in Figure 2, we formalize a structured scenario space defined by two orthogonal axes, namely *product*

inventory and *customer personas*, using script template provided in details in Appendix H. As shown in raw data stage and item generation stage in Figure 2, we instantiate this space by (i) synthesizing and deduplicating product inventories for online consumer goods and retail financial services (see Appendix B.1 for detailed statistics). Specifically, for financial services, we collect 300 real-world products across different categories and randomly sample 10 examples as few-shot demonstrations. For consumer goods, we leverage the Amazon-Reviews-2023 dataset (Hou et al., 2024), sampling one target item and five related product descriptions from the corresponding category. In both cases, we guide the LLM to generate product profiles enriched with detailed selling points. (ii) constructing a pool of base customer personas by sampling core attributes including age, occupation, and city, which serves as the foundation for generating more enriched, product-conditioned persona profiles in the subsequent script instantiation stage. The resulting Cartesian product defines the scenario space used for script instantiation as in scripts instantiation stage in Figure 2. The distribution of decision factors across age groups is visualized in Figure 3.

Table 1: Difficulty profiles used to control buy propensity and buyer style for the simulated customer.

Profile	Buy prop. p_k	Buyer style
<i>easy</i>	0.80	Open-minded, motivated buyer with clear pain points and flexible budget; decides quickly if the product is a plausible fit.
<i>medium</i>	0.50	Balanced buyer with concrete but resolvable concerns (e.g., price or fit); requires reasonable evidence and engages in moderate objection handling.
<i>hard</i>	0.20	Skeptical, price-sensitive, and risk-averse buyer; defaults to negative purchase intent unless strong, specific evidence and clear ROI are demonstrated.
<i>very hard</i>	0.10	Highly skeptical enterprise buyer with strict compliance and procurement constraints; requires detailed proof, references, and process alignment, typically postponing purchase.
<i>adversarial</i>	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.

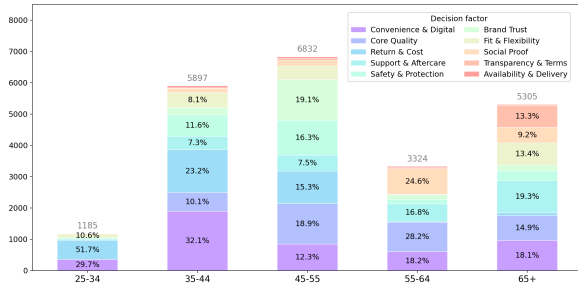


Figure 3: Decision Factor Distribution by Age Group. Stacked bars show the total number of decision-factor mentions (*Count*) for each age group in the SalesLLM persona set. Colors denote decision-factor categories (Return & Cost, Convenience & Digital, Fit & Flexibility, Safety & Protection, Support & 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.

Detailed definitions of these decision factors, along with comprehensive statistics on product inventories and persona distributions, are provided in Appendix B.

Controllable difficulty and intent. We model scenario difficulty with five calibrated customer profiles (*easy*, *medium*, *hard*, *very hard*, *adversarial*). Each profile k defines two interpretable controls (Table 1): (i) prior buy propensity $p_k \in [0, 1]$ (0.8 for *easy* to 0.05 for *adversarial*); (ii) a canonical buyer style from cooperative to adversarial. 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 multi-turn agent benchmarks (Ma et al., 2024; Guan et al., 2024).

Standardized script set. To ensure that simulated customers exhibit realistic pain points and purchase motivations aligned with specific products—thereby enabling the user model to accurately simulate internal psychological transitions during the dialogue—we construct a large-scale dataset of 30,074 dialogue scripts. We formalize the script generation as a hierarchical conditional sampling process. First, a base scenario configuration is sampled from the joint distribution of products P , base personas U (characterized by age, occupation, and city), and difficulty profiles K :

$$(p, u, k) \sim p(P, U, K). \quad (1)$$

To ensure the realism of the interaction, we then sample enriched persona attributes a —including product-conditioned *motivations*, *decision factor*, *pain points*, and *initial utterances*—conditioned on the selected product and base persona:

$$a \sim p(A | p, u). \quad (2)$$

Finally, the script generator G produces the natural-language specifications for both the sales agent and the user simulator based on the complete scenario state $s = (p, u, a, k)$:

$$(\text{script}^{\text{sales}}, \text{script}^{\text{user}}) \sim p(G | s). \quad (3)$$

This sampling process is implemented using Qwen3-72B (Yang et al., 2025), with MinHash-based deduplication (Broder, 1997) applied to maintain high dataset diversity.

For evaluation, we select 1,000 Chinese scripts and create a parallel English subset by translating them and manually filtering low-quality translations, resulting in 805 English scripts.

3.2 User-based Model

To further improve the dialogue realism of SalesLLM, our goal is to post-train user-based models to simulate users that can be conditioned on

a sales task and a user intent to replicate human conversational behavior, and achieve interaction with sales assistants to steer conversations towards sales tasks we want to study. To archive realistic user simulation for SalesLLM, however, we need to resolve two challenges we found during our experiments, namely, **language bias** and **role consistency**.

Language Bias. A key limitation of generic LLM-based simulators is their tendency to produce stylistically rigid outputs. Synthetic user utterances often exhibit patterns that diverge from human behavior: excessive explicitness, repetitive structures, and reduced interactional variability. These artifacts reduce the ecological validity of simulated dialogues and limit their utility for evaluating persuasion strategies.

Role Consistency. Another key challenge in LLM-based user simulation is that common LLMs are primarily trained to respond helpfully to user queries. When repurposed for generating customer utterances, these models often drift toward their assistant-style behavior, producing responses that resemble helpful answers rather than authentic customer communication. This *role inversion* becomes more pronounced in multi-turn dialogues, compromising simulation reliability.

To solve those two challenges, in this section, we introduce our user-based model, CustomerLM, a specialized simulator designed to generate realistic customer behavior in sales dialogues. Related user simulators have explored similar directions (Wang et al., 2025; Naous et al., 2025), but often rely on synthetic or open-domain data. To reduce language bias, we train on a human-annotated dataset of customer responses collected from real-world marketing interactions, which contains natural discourse markers and highly variable, sometimes fragmented utterances. To improve role consistency, we apply a two-stage SFT (Brown et al., 2020) followed by DPO (Rafailov et al., 2023); comparisons with GPT-4o, UserLM (Naous et al., 2025), and USP (Wang et al., 2025) are reported in Section 4.4.

We build our user model, CustomerLM, upon Qwen3-8B (Yang et al., 2025), and train it on 8,000 crowd worker-involved real-world sales dialogues (The statical charts of the data are shown in Table 2), where each of the 80 crowd workers contributed 100 interaction sessions with an average of 3.74 turns. As shown in Figure 4, We employ a two-stage training pipeline: First, SFT is applied to

learn authentic customer response patterns and conversational tone from real-world sales dialogues. Second, we construct a preference dataset for DPO by extracting 1,000 script pairs from held-out evaluation scenarios and generating multi-turn dialogues with various LLMs serving as the sales agent, including GPT-4o, GLM-4.6 (Zhipu AI (Z.ai), 2025), and Qwen3-72B (Alibaba Cloud, 2024), totaling 3,000 dialogues. We then use LLM-as-a-judge to select contrastive examples, followed by manual filtering to curate 268 pairs. Positive examples are augmented with optimal user responses generated by GPT-4o to form the final DPO training set. As demonstrated in our ablation study (Section 4.4), this custom model achieves substantially better dialogue quality and role consistency compared to the comparison user simulators.

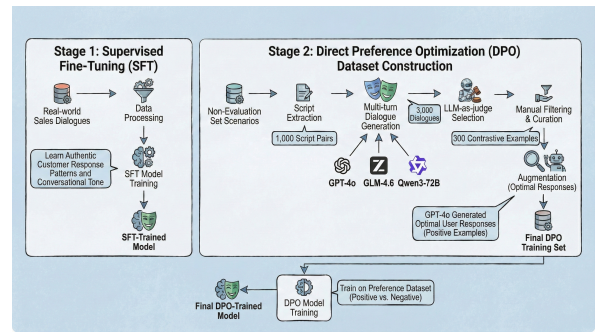


Figure 4: SalesLLM is trained in two stages: first, SFT on 8,000 crowd-worker sales dialogues to learn realistic customer responses; second, DPO using 300 manually curated contrastive dialogue pairs, augmented with LLM-generated optimal user responses.

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 2: The statical 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.

3.3 Scoring Methodology

Evaluating open-ended dialogue systems remains challenging, with recent work increasingly adopting LLM-based judges for multi-dimensional assessment (Zheng et al., 2023; Zhou et al., 2024; Chen et al., 2024). Unlike rule-based metrics, LLM judges leverage superior reasoning capabilities to evaluate complex interaction quality. However,

316	domain-specific applications demand tailored evaluation frameworks that capture unique success criteria. In sales dialogues, effectiveness is reflected in two orthogonal dimensions: (1) the customer’s purchase intent at conversation end, and (2) the salesman’s ability to advance the deal through strategic communication. While buying intention serves as a direct indicator of sales success (Islam et al., 2011), it alone cannot fully capture the assistant’s selling proficiency—effective persuasion may cultivate long-term customer relationships even without immediate purchase. We therefore design a dual-metric scoring system that jointly assesses buying intention and selling performance, enabling comprehensive evaluation of sales effectiveness.	366
317		367
318		368
319		369
320		370
321		371
322		372
323		373
324		374
325		375
326		376
327		377
328		378
329		379
330		380
331	3.3.1 Scoring User Buying Intention	381
332	We assume sales dialogues have relatively clear signals for the user’s buying intentions, such as when users say “I don’t want it,” and these signals (such as decision messages) are commonly located at the end of the dialogues. To score the users’ buying intention, we train two BERT (Devlin et al., 2019) models (one for Chinese, chinese-roberta-wwm-ext (Cui et al., 2020), and one for English, roberta-base (Liu et al., 2019)) as buying intention scorers. BERT-like models are more suitable for texts that have surface-level patterns than other LLM-based models (Zhang et al., 2025a), making them well-suited for detecting explicit buying intention signals in dialogue text.	382
333		383
334		384
335		385
336		386
337		387
338		388
339		389
340		390
341		391
342		392
343		393
344		394
345		395
346		396
347		397
348		398
349		399
350		400
351		401
352		402
353		403
354		404
355		405
356		406
357	3.3.2 Scoring Salesmen Selling Performance	407
358	Evaluating selling performance requires grounding in established psychological and communication theories. Drawing on sales psychology (Tracy, 2006), which identifies need identification, objection handling, and commitment elicitation as core selling competencies, persuasive communication research (Taillard, 2000), which emphasizes observable behavioral outcomes as persuasion indicators, and adaptive selling theory (Weitz, 1978), which links salesperson understanding of customer decision-making to performance outcomes, we operationalize selling ability through four core dimensions: (1) verbal purchase commitment, (2) concrete next-step alignment (e.g., scheduling demos, locking evaluation milestones), (3) key information elicitation (budget, decision-makers, timeline), and (4) objection resolution effectiveness. As detailed in Appendix G, the judge assesses dialogue progression signals embedded in user responses and outputs a 0–10 score, where higher scores reflect stronger evidence of intent escalation and actionable forward progress. This approach follows dialogue quality assessment protocols (Juster, 1966) and leverages LLM reasoning to capture nuanced interaction patterns that rule-based metrics cannot fully model.	408
359		409
360		410
361		411
362		412
363		413
364		414
365		415
	3.3.3 Final Score	416
	Finally, as the final score, we combine the buying intention score and the selling performance score with weights as shown in Equation 4:	417
		418
	$\text{final_score} = \alpha \cdot \text{buying} + (1 - \alpha) \cdot \text{selling} \quad (4)$	419
		420
		421
		422
		423
		424
		425
		426
		427
		428
		429
		430
		431
		432
		433
		434
		435
		436
		437
		438
		439
		440
		441
		442
		443
		444
		445
		446
		447
		448
		449
		450
		451
		452
		453
		454
		455
		456
		457
		458
		459
		460
		461
		462
		463
		464
		465
		466
		467
		468
		469
		470
		471
		472
		473
		474
		475
		476
		477
		478
		479
		480
		481
		482
		483
		484
		485
		486
		487
		488
		489
		490
		491
		492
		493
		494
		495
		496
		497
		498
		499
		500
		501
		502
		503
		504
		505
		506
		507
		508
		509
		510
		511
		512
		513
		514
		515
		516
		517
		518
		519
		520
		521
		522
		523
		524
		525
		526
		527
		528
		529
		530
		531
		532
		533
		534
		535
		536
		537
		538
		539
		540
		541
		542
		543
		544
		545
		546
		547
		548
		549
		550
		551
		552
		553
		554
		555
		556
		557
		558
		559
		560
		561
		562
		563
		564
		565
		566
		567
		568
		569
		570
		571
		572
		573
		574
		575
		576
		577
		578
		579
		580
		581
		582
		583
		584
		585
		586
		587
		588
		589
		590
		591
		592
		593
		594
		595
		596
		597
		598
		599
		600
		601
		602
		603
		604
		605
		606
		607
		608
		609
		610
		611
		612
		613
		614
		615
		616
		617
		618
		619
		620
		621
		622
		623
		624
		625
		626
		627
		628
		629
		630
		631
		632
		633
		634
		635
		636
		637
		638
		639
		640
		641
		642
		643
		644
		645
		646
		647
		648
		649
		650
		651
		652
		653
		654
		655
		656
		657
		658
		659
		660
		661
		662
		663
		664
		665
		666
		667
		668
		669
		670
		671
		672
		673
		674
		675
		676
		677
		678
		679
		680
		681
		682
		683
		684
		685
		686
		687
		688
		689
		690
		691
		692
		693
		694
		695
		696
		697
		698
		699
		700
		701
		702
		703
		704
		705
		706
		707
		708
		709
		710
		711
		712
		713
		714
		715
		716
		717
		718
		719
		720
		721
		722
		723
		724
		725
		726
		727
		728
		729
		730
		731
		732
		733
		734
		735
		736
		737
		738
		739
		740
		741
		742
		743
		744
		745
		746
		747
		748
		749
		750
		751
		752
		753
		754
		755
		756
		757
		758
		759
		760
		761
		762
		763
		764
		765
		766
		767
		768
		769
		770
		771
		772
		773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
		790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800
		801
		802
		803
		804
		805
		806
		807

4 LLMs Results on SalesLLM

In this section, we present experimental results evaluating multiple mainstream LLMs on SalesLLM. 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 14 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 (ByteDance, 2025), Qwen3-max (Alibaba Cloud, 2024), Deepseek-chat (AI, 2025), GLM-4.6 (Zhipu AI (Z.ai), 2025), GPT-4o, GPT-5-nano (OpenAI, 2025), Gemini-3-Flash (DeepMind, 2025), and Xiaomi-MiMo-V2 (Xiaomi, 2025). For open-weights models, we evaluate Llama-3.3-70B (Grattafiori et al., 2024), Qwen3 (8B/32B/72B) (Yang et al., 2025), and Gemma-3-27B (Team et al., 2025). 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 8 for human salesmen to talk to the user who is played by GPT-4o.

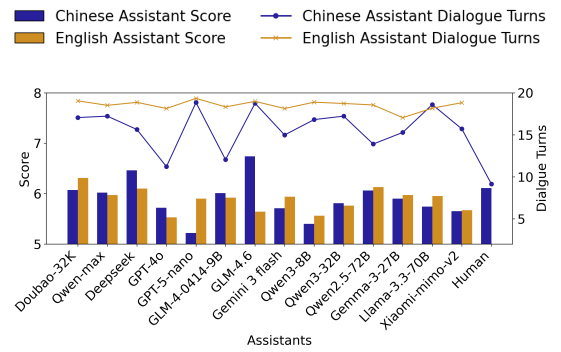
4.2 Overall Performance

Table 10 and Figure 5 presents the overall performance of different models. Results show significant variation in selling abilities, with stronger models achieving consistently higher SalesLLM scores. Notably, top-tier models like Doubao-1.5 and GLM-4.6 outperform the human baseline (Human Salesmen Group) in Chinese scenarios, highlighting the potential of LLMs in sales domains.

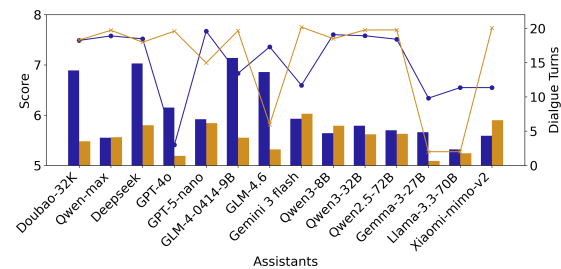
While models like Doubao-1.5 excel in Chinese, they show significant performance drops in English when interacting with one version of our CustomerLM User model (6.89 \rightarrow 5.48). In contrast, Gemini-3 demonstrates robust cross-lingual consistency (6.39/6.03). Qualitative analysis (Appendix I) 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 Figure 5, 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 SalesLLMs to future work.



(a) GPT-4o as User Model



(b) CustomerLM as User Model

Figure 5: Distribution of SalesLLM scores vs. dialogue turns across 1,805 scenarios (Chinese and English).

4.3 Human Evaluation Validation

To validate our automated scoring methodology, we compare SalesLLM scores against 8 human evaluations on 100 randomly sampled dialogues. We implement the interface (shown in Figure 9) for 8 human scorers. As shown in Figure 6, the automated scores demonstrate strong alignment with human judgments. Specifically, we achieve a Pearson correlation coefficient of $r = 0.98$ (Pearson, 1895) and a Spearman's rank correlation of $\rho = 1.0$ (Spearman, 1904). These results significantly exceed the common threshold for high correlation ($r > 0.6$) (Cohen, 1988), confirming the reliability of SalesLLM score as an automated evaluation metric.

Table 3: Ablation study comparing our CustomerLM as user model against GPT-4o, UserLM and USP as user models. Higher is better for dialogue quality metrics; lower is better for role inversion rates.

User Model	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	Sem. Sim.	Role Inversion (%)
GPT-4o	0.10	0.08	0.02	0.07	0.57	17.44
UserLM	0.06	0.08	0.01	0.06	0.50	21.55
USP	0.08	0.09	0.01	0.08	0.52	18.76
CustomerLM (ours)	0.12	0.11	0.03	0.10	0.59	8.8

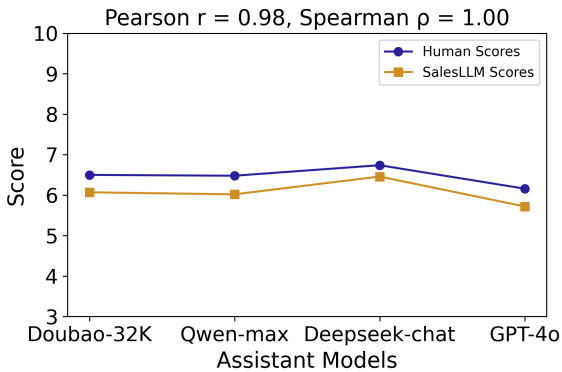


Figure 6: Correlation analysis between Human evaluation scores and SalesLLM scores on 100 randomly selected dialogues.

4.4 Ablation Study of Custom User Model

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 5.

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-1/2/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.

Results. Table 3 shows Our model outperforms comparison models on turn-level similarity (BLEU-4, ROUGE-1/2/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, a comprehensive benchmark for evaluating the realistic selling capabilities of Large Language Models. By formalizing a structured scenario space of 30,074 potential scripts and instantiating a diverse, difficulty-calibrated test set of 1,805 scenarios, SalesLLM enables systematic assessment of sales performance across varying conditions. We also introduced a custom user-based model, CustomerLM, trained on human-salesmen dialogues, which significantly improves simulation realism by reducing role inversion and enhancing stylistic alignment with authentic customer behavior. Our dual-metric evaluation framework, validated against human judgment ($r = 0.98$), provides a robust standard for measuring both purchase intent and selling proficiency. Extensive experiments reveal that while top-tier LLMs demonstrate promising sales potential—surpassing human baselines in specific Chinese scenarios—significant gaps remain in cross-lingual consistency and complex persuasion. The potential future works should be further researching on the language impacts for SalesLLM; including more complex selling scenarios in SalesLLM (such as long-term-memory-based relationship management); and etc. We hope SalesLLM serves as a foundational platform for advancing autonomous sales agents and broader goal-oriented dialogue systems.

550 Limitations

551 We identify three primary limitations. First, while
552 we employ human calibration, the LLM-based user
553 simulator may not fully capture the complexity of
554 real customer behavior, especially regarding emo-
555 tional fluctuations and trust building in long-term
556 dialogues (de Wit, 2023; Sekulic et al., 2024; Wang
557 et al., 2025). This simulation gap is consistent with
558 findings in other recent studies (Sun et al., 2024;
559 Luo et al., 2024; Algherairy and Ahmed, 2025).
560 Second, SalesLLM currently models single-session
561 interactions. Real-world sales cycles often span
562 multiple exchanges requiring long-term memory
563 and relationship management, which remain fu-
564 ture work. Third, we observe that agents occasion-
565 ally hallucinate unauthorized concessions—such
566 as discounts absent from the script—to maximize
567 conversion rates, despite explicit instructions to
568 the contrary. This unfaithful persuasion poses a
569 challenge for reliable evaluation, echoing validity
570 concerns in broader LLM benchmarks (Li et al.,
571 2024).

572 References

- 573 DeepSeek AI. 2025. Deepseek chat. Large lan-
574 guage model, online AI chatbot. [https://chat.
575 deepseek.com/](https://chat.deepseek.com/).
- 576 Atheer Algherairy and Moataz Ahmed. 2025. [Prompt-](#)
577 [ing large language models for user simulation in task-](#)
578 [oriented dialogue systems.](#) *Comput. Speech Lang.*,
579 89(C).
- 580 Alibaba Cloud. 2024. Qwen-max. Large language
581 model. <https://qwenlm.github.io/>.
- 582 A. Broder. 1997. On the resemblance and containment
583 of documents. In *Proceedings of the Compression
584 and Complexity of Sequences 1997*, SEQUENCES
585 '97, page 21, USA. IEEE Computer Society.
- 586 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie
587 Subbiah, Jared D. Kaplan, Prafulla Dhariwal,
588 Arvind Neelakantan, Pranav Shyam, Girish Sastry,
589 Amanda Askell, Sandhini Agarwal, Ariel Herbert-
590 Voss, Gretchen Krueger, Tom Henighan, Rewon
591 Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey
592 Wu, Clemens Winter, and 12 others. 2020. Lan-
593 guage models are few-shot learners. *arXiv preprint
594 arXiv:2005.14165*.
- 595 ByteDance. 2025. Doubao-1.5-pro-32k. Large lan-
596 guage model, Volcano Engine. [https://www.
597 volcengine.com/product/doubao](https://www.volcengine.com/product/doubao).
- 598 Wen-Yu Chang and Yun-Nung Chen. 2024. [Injecting](#)
599 [salesperson’s dialogue strategies in large language](#)

models with chain-of-thought reasoning. *Preprint*,
arXiv:2404.18564. 600 601

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang
Tang. 2017. [A survey on dialogue systems: Recent
advances and new frontiers.](#) *SIGKDD Explor. Newsl.*,
19(2):25–35. 602 603 604 605

Hongzhan Chen, Hehong Chen, Ming Yan, Wen-
shen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan,
Chenliang Li, Ji Zhang, Fei Huang, and Jingren
Zhou. 2024. [Socialbench: Sociality evaluation
of role-playing conversational agents.](#) *Preprint*,
arXiv:2403.13679. 606 607 608 609 610 611

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu
Lian, and Zheng Liu. 2025. [M3-embedding: Multi-
linguality, multi-functionality, multi-granularity text
embeddings through self-knowledge distillation.](#) *Preprint*, arXiv:2402.03216. 612 613 614 615 616

Jie Chen, Wei Liu, Li Zhang, Hao Wang, and Ming
Zhao. 2026. [Breaking single-tester limits: Multi-
agent LLMs for multi-user feature testing.](#) In *Pro-
ceedings of the 2026 IEEE/ACM 48th International
Conference on Software Engineering*, pages 1–12.
ACM. 617 618 619 620 621 622

Sijia Cheng, Wen Yu Chang, and Yun-Nung Chen. 2025.
[Exploring personality-aware interactions in salesper-
son dialogue agents.](#) In *Proceedings of the 15th In-
ternational Workshop on Spoken Dialogue Systems
Technology*, pages 60–71, Bilbao, Spain. Association
for Computational Linguistics. 623 624 625 626 627 628

Jacob Cohen. 1988. *Statistical power analysis for the
behavioral sciences.* Routledge. 629 630

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 Computa-
tional Linguistics. 631 632 633 634 635 636 637

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, CONVERSA-
TIONS 2023, Oslo, Norway, November 22–23, 2023*,
Revised Selected Papers, page 77–93, Berlin, Heidel-
berg. Springer-Verlag. 638 639 640 641 642 643 644

DeepMind. 2025. Gemini 3 flash model evalua-
tion. [https://storage.googleapis.com/
deepmind-media/gemini/gemini_3_flash_
model_evaluation.pdf](https://storage.googleapis.com/deepmind-media/gemini/gemini_3_flash_model_evaluation.pdf). Accessed: 2025-12-28. 645 646 647 648

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and
Kristina Toutanova. 2019. BERT: Pre-training of
deep bidirectional transformers for language under-
standing. In *Proceedings of the 2019 Conference of
the North American Chapter of the Association for*
649 650 651 652 653

654		Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	710
655			711
656			712
657			713
658	Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators . <i>Preprint</i> , arXiv:2404.04475.		714
659			715
660			716
661			717
662	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 . <i>Preprint</i> , arXiv:2407.21783.		718
663			719
664			720
665			721
666			722
667			723
668			724
669			725
670	Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy . <i>Preprint</i> , arXiv:2407.06813.		726
671			727
672			728
673			729
674	Leon Hanschmann, Ulrich Gnewuch, and Alexander Maedche. 2023. Saleshat: A LLM-based social robot for human-like sales conversations . pages 61–76, Berlin, Heidelberg. Springer-Verlag.		730
675			731
676			732
677			733
678	Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiushi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation . <i>Preprint</i> , arXiv:2403.03952.		734
679			735
680			736
681			737
682	Md Aminul Islam, Ku Amir Ku Daud, and 1 others. 2011. Factors that influence customers’ buying intention on shopping online. <i>International Journal of marketing studies</i> , 3(1):128.		738
683			739
684			740
685			741
686	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 <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.		742
687			743
688			744
689			745
690			746
691			747
692			748
693			749
694	F. Thomas Juster. 1966. Consumer buying intentions and purchase probability: An experiment in survey design . <i>Journal of the American Statistical Association</i> , 61(315):658–696.		750
695			751
696			752
697			753
698	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 <i>Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles</i> .		754
699			755
700			756
701			757
702			758
703			759
704			760
705	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 benchmark pipeline . <i>CoRR</i> , abs/2406.11939.		761
706			762
707			763
708			764
709			765
			766
			767
			768
			769
			770
			771
			772
			773
			774
			775
			776
			777
			778
			779
			780
			781
			782
			783
			784
			785
			786
			787
			788
			789
			790
			791
			792
			793
			794
			795
			796
			797
			798
			799
			800
			801
			802
			803
			804
			805
			806
			807
			808
			809
			810
			811
			812
			813
			814
			815
			816
			817
			818
			819
			820
			821
			822
			823
			824
			825
			826
			827
			828
			829
			830
			831
			832
			833
			834
			835
			836
			837
			838
			839
			840
			841
			842
			843
			844
			845
			846
			847
			848
			849
			850
			851
			852
			853
			854
			855
			856
			857
			858
			859
			860
			861
			862
			863
			864
			865
			866
			867
			868
			869
			870
			871
			872
			873
			874
			875
			876
			877
			878
			879
			880
			881
			882
			883
			884
			885
			886
			887
			888
			889
			890
			891
			892
			893
			894
			895
			896
			897
			898
			899
			900
			901
			902
			903
			904
			905
			906
			907
			908
			909
			910
			911
			912
			913
			914
			915
			916
			917
			918
			919
			920
			921
			922
			923
			924
			925
			926
			927
			928
			929
			930
			931
			932
			933
			934
			935
			936
			937
			938
			939
			940
			941
			942
			943
			944
			945
			946
			947
			948
			949
			950
			951
			952
			953
			954
			955
			956
			957
			958
			959
			960
			961
			962
			963
			964
			965
			966
			967
			968
			969
			970
			971
			972
			973
			974
			975
			976
			977
			978
			979
			980
			981
			982
			983
			984
			985
			986
			987
			988
			989
			990
			991
			992
			993
			994
			995
			996
			997
			998
			999
			1000

767	OpenAI. 2025. Gpt-5 system card. https://openai.com/index/gpt-5-system-card . Accessed: 2025-12-28.	820
768		821
769		822
770	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	823
771	Jing Zhu. 2002. Bleu: a method for automatic evalu-	824
772	ation of machine translation. In <i>Proceedings of the</i>	
773	<i>40th Annual Meeting of the Association for Comput-</i>	
774	<i>ational Linguistics</i> , pages 311–318, Philadelphia,	
775	Pennsylvania, USA. Association for Computational	
776	Linguistics.	
777	Karl Pearson. 1895. Note on regression and inheritance	
778	in the case of two parents. <i>Proceedings of the Royal</i>	
779	<i>Society of London</i> , 58:240–242.	
780	Richard E. Petty and John T. Cacioppo. 1986. <i>Com-</i>	
781	<i>munication and Persuasion: Central and Peripheral</i>	
782	<i>Routes to Attitude Change</i> . Springer-Verlag.	
783	Rafael Rafailov, Archit Sharma, Eric Mitchell, Ste-	
784	fano Ermon, Christopher D. Manning, and Chelsea	
785	Finn. 2023. Direct preference optimization: Your	
786	language model is secretly a reward model. <i>ArXiv</i> ,	
787	abs/2305.18290.	
788	Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin,	
789	and Sameer Singh. 2020. Beyond accuracy: Be-	
790	havioral testing of NLP models with CheckList. In	
791	<i>Proceedings of the 58th Annual Meeting of the Asso-</i>	
792	<i>ciation for Computational Linguistics</i> , pages 4902–	
793	4912, Online. Association for Computational Lin-	
794	guistics.	
795	Ananya B. Sai, Akash Kumar Mohankumar, and	
796	Mitesh M. Khapra. 2020. A survey of evalu-	
797	ation metrics used for nlg systems. <i>Preprint</i> ,	
798	arXiv:2008.12009.	
799	Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and	
800	Steve Young. 2006. A survey of statistical user sim-	
801	ulation techniques for reinforcement-learning of di-	
802	alogue management strategies. <i>Knowl. Eng. Rev.</i> ,	
803	21(2):97–126.	
804	Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia	
805	Khau, Bruna Guedes, Modestas Filipavicius, An-	
806	dre Ferreira Manso, and Roland Mathis. 2024. Re-	
807	liable LLM-based user simulator for task-oriented	
808	dialogue systems. In <i>Proceedings of the 1st Work-</i>	
809	<i>shop on Simulating Conversational Intelligence in</i>	
810	<i>Chat (SCI-CHAT 2024)</i> , pages 19–35, St. Julians,	
811	Malta. Association for Computational Linguistics.	
812	Charles Spearman. 1904. The proof and measurement	
813	of association between two things. <i>American Journal</i>	
814	<i>of Psychology</i> , 15:72–101.	
815	Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller,	
816	Becka Silvert, Bing Liu, Zhiguang Wang, Hon-	
817	glei Liu, Eunjoon Cho, and Claire Cardie. 2021.	
818	Adding chit-chat to enhance task-oriented dialogues.	
819	<i>Preprint</i> , arXiv:2010.12757.	
	Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren,	825
	Zhumin Chen, Maarten de Rijke, and Zhaochun Ren.	826
	2024. Metaphorical user simulators for evaluating	827
	task-oriented dialogue systems. <i>ACM Transactions</i>	
	<i>on Information Systems</i> , 42(1):17:1–17:29.	
	Marie-Odile Taillard. 2000. Persuasive communication:	
	the case of marketing. <i>Working Papers in Linguistics</i> ,	
	12:145–174.	
	Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya	828
	Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin,	829
	Tatiana Matejovicova, Alexandre Ramé, Morgane	830
	Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey	831
	Cideron, Jean bastien Grill, Sabela Ramos, Edouard	832
	Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev,	833
	and 197 others. 2025. Gemma 3 technical report.	834
	<i>Preprint</i> , arXiv:2503.19786.	835
	Brian Tracy. 2006. <i>The psychology of selling: Increase</i>	836
	<i>your sales faster and easier than you ever thought</i>	837
	<i>possible</i> . HarperCollins Leadership.	838
	Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou,	839
	Feng Jiang, and Haizhou Li. 2025. Know you first	840
	and be you better: Modeling human-like user sim-	841
	ulators via implicit profiles. In <i>Proceedings of the</i>	842
	<i>63rd Annual Meeting of the Association for Comput-</i>	843
	<i>ational Linguistics (Volume 1: Long Papers)</i> , pages	844
	21082–21107, Vienna, Austria. Association for Com-	845
	putational Linguistics.	846
	Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung	847
	Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2020.	848
	Persuasion for good: Towards a personalized per-	849
	suasive dialogue system for social good. <i>Preprint</i> ,	850
	arXiv:1906.06725.	851
	Zekun Moore Wang, Zhongyuan Peng, Haoran Que,	852
	Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu,	853
	Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian	854
	Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang,	855
	Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng.	856
	2024. Rolellm: Benchmarking, eliciting, and enhanc-	857
	ing role-playing abilities of large language models.	858
	<i>Preprint</i> , arXiv:2310.00746.	859
	Barton A Weitz. 1978. Relationship between sales-	860
	person performance and understanding of customer	861
	decision making. <i>Journal of Marketing Research</i> ,	862
	15(4):501–516.	863
	LLM-Core Xiaomi. 2025. Mimo-v2-flash technical	864
	report.	865
	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,	866
	Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,	867
	Chengen Huang, Chenxu Lv, Chujie Zheng, Day-	868
	iheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao	869
	Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41	870
	others. 2025. Qwen3 technical report. <i>ArXiv</i> ,	871
	abs/2505.09388.	872
	Junyan Zhang, Yiming Huang, Shuliang Liu, Yubo Gao,	873
	and Xuming Hu. 2025a. Do bert-like bidirectional	874
	models still perform better on text classification in	875
	the era of llms? <i>arXiv preprint arXiv:2505.18215</i> .	876

877 Qingyu Zhang, Chunlei Xin, Xuanang Chen, Yaojie Lu,
878 Hongyu Lin, Xianpei Han, Le Sun, Qing Ye, Qian-
879 long Xie, and Xingxing Wang. 2025b. *Ai-salesman:*
880 *Towards reliable large language model driven tele-*
881 *marketing. Preprint*, arXiv:2511.12133.

882 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan
883 Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,
884 Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,
885 Joseph E. Gonzalez, and Ion Stoica. 2023. *Judg-*
886 *ing llm-as-a-judge with mt-bench and chatbot arena.*
887 *Preprint*, arXiv:2306.05685.

888 Zhipu AI (Z.ai). 2025. Glm-4-6b. Open-source
889 large language model. [https://huggingface.co/](https://huggingface.co/THUDM/GLM-4-6B)
890 [THUDM/GLM-4-6B](https://huggingface.co/THUDM/GLM-4-6B).

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

A User-Model Evaluation Details

This section details the automatic evaluation protocol for the user simulator described in Section 4.4.

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 4: The static 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.

Below are the system prompt used with CustomerLM as the user model, along with one example of test data.

System Prompt

Your task:

You are a customer management expert, and you need to refer to the following content for communication and response. The customer is a customer service representative from xx Securities, a company formerly known as xx Securities. They are promoting a special offer for existing customers, which provides a stable annual return of 0.035. Specifically, an investment of 50,000 yuan over three years can yield a return of 5,250 yuan, with funds supervised by a bank to ensure safety. The customer also suggests adding the company’s WeChat account for more information and assistance.

Rules:

- Use English strictly for the conversation.
- Be realistic and rational; do not mention things that are impossible.
- Avoid discussing personal life or other topics

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 5: Held-out ablation test set statistics (Sec. 4.4). Averages exclude the system prompt. Lang is language, Dia is dialogues.

unrelated to your purchase intention.
-Use conversational language instead of formal writing; keep each response brief.
-Do not fabricate product facts; respond from a customer perspective based on your understanding.
-Avoid using parentheses to describe your actions or inner thoughts; just speak naturally.

One example of test data

assistant: Hi.

user: Hello, I am a customer service representative from xx Securities. Our company’s previous name was xx Securities, do you remember?

assistant: What?

user: The company formerly known as xx Securities has now been renamed to xx Securities. I remember you have used our services before, right?

assistant: Oh, I remember now.

user: That’s correct, the original XX Securities is now called XX Securities. We are reaching out to you today because there is a special promotional event for our long-time clients, where you can receive a stable annual return of 0.035. May I provide you with more details?

assistant: Okay, please tell me the details.

user: Of course. This activity is very simple: you just need to join with your idle funds and become our esteemed member to receive corresponding returns. For example, if you invest 50,000 yuan for three years, you can receive a total return of 5,250 yuan, with an annualized return rate of 0.035. These returns can be directly transferred to your bank account. Additionally, the funds are regulated by the bank to ensure the safety of your principal. Does this explanation help you understand?

assistant: I understand, but I still have some concerns about security.

user: Rest assured, your funds will be deposited in your personal bank account and monitored by the bank. It's like a margin, providing you with membership privileges and vehicle insurance protection through our platform.

assistant: Okay.

user: If this activity interests you, you can open WeChat and follow our official account. Are you able to use WeChat now? You can start by adding our corporate WeChat account.

assistant: I don't have time right now, but I have already learned some information.

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.1. 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 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

Please analyze the following sales conversation and determine if the USER incorrectly acted as the ASSISTANT.

Conversation: {conversation_text}

Your Task: Check the USER's messages to see if they exhibited "Role Reversal" (i.e., the USER started pitching products, offering services, or acting like a customer service agent/salesperson).

Criteria: 1. **Role Reversal:** The USER proactively explains product features, offers quotes, appeases the customer, or asks "How can I help you?" - behaviors clearly belonging to a sales/support role. 2. **Normal Interaction:** The USER asks questions, states needs, bargains, refuses, or accepts. These are normal USER behaviors.

Please output JSON in this format: { "detected": true/false, "severity": "none/low/medium/high", "reason": "Brief reason for judgment", "examples": ["Incorrect statement by USER 1", "Incorrect statement by USER 2"] }

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

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

Table 7: Statistics of the SalesLLM scenario space.

Note: - Set 'detected' to true if ANY obvious sales/support behavior is found in USER messages. - Output ONLY JSON. Do not include Markdown formatting or other text.

A.2 Results

Table 6 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.

B Scenario and Persona Statistics

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

B.1 Scenario Space Overview

Table 7 summarizes the scale of our scenario components. We synthesized 10,074 E-commerce items from 33 seed categories (Table 8) and 20,000 financial services products (Table 9) 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.

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 8: Category distribution of the E-commerce scenario space in SalesLLM.

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.

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 9: Distribution of Financial Services categories.

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.

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%
- **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

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 hand-held.
 - **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 Results of the LLM Evaluations on SalesLLM

The number results of LLM evaluations.

Table 10: Overall performance (SalesLLM Score) of different Assistant Models when evaluated against two user simulators: GPT-4o and our Custom User 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 †.

Assistant Model	GPT-4o User		Custom User	
	ZH	EN	ZH	EN
Doubao-1.5-pro-32k	6.50	<u>6.31</u>	6.89 [†]	5.48
Qwen-Max	6.48	5.97	5.55	5.56
DeepSeek-Chat (V3.1)	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	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
Qwen3-72B	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	–	–	–

E BERT vs. GPT-4o Classifier

We validate our fine-tuned BERT classifier against GPT-4o (LLM-as-a-Judge). Table 11 shows that our specialized model achieves significantly higher accuracy (>90%) compared to GPT-4o (68 – 74%) for buying intention detection.

F Script Examples

Example of a customer script and a salesperson script.

User Script Example

- Difficulty level: medium
- Buy-inclination score: 0.6
- Persona: Balanced buyer. Has some concerns

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 11: 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.

(price/fit) and needs reasonable evidence. Will buy if key objections are addressed satisfactorily. Inclined to purchase, but will raise reasonable questions and concerns. If they are confident enough about the product and these issues are resolved satisfactorily, they may buy.

- CUSTOMER_INFORMATION (private to you):
- basic_info: "age_group": "45-55", "gender": "male", "location": "Shanghai", "occupation": "senior business consultant"
- motivations: Hopes to subscribe to a professional technology magazine to obtain cutting-edge technical insights, innovation trends, and cultural dynamics to assist their business consulting work and fulfill their interest in the technology field.
- pain_points: Has doubts about product quality authenticity, worries about inadequate after-sales service and payment security issues; at the same time hopes that the subscription magazine's content is professionally detailed and has good brand credibility.
- decision_factors: Values brand credibility, product reviews, after-sales guarantees, payment security, and logistics speed; also pays attention to the professionalism, depth, and uniqueness of the magazine content.
- communication_preferences: Prefers WeChat and email; likes formal, professional, detailed content with text and images.
- language: english

Salesperson Script Example

You are a professional salesperson (ASSISTANT) in a realistic sales conversation. Only you can see the following PRODUCT_INFORMATION. Never reveal it or where it came from. Speak naturally and be helpful.

You should use sales strategies and techniques to persuade the customer. If the customer has committed to buying the product, you should provide a purchase channel if it is available in PRODUCT_INFORMATION. else, you should provide a reason why the customer should buy the product.

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
 store: TechMagStore
 categories:

Magazine Subscriptions
 Technology
 Computers & Electronics
 details:
 Subscription Type: Print Only
 Frequency: Monthly
 Number of Issues: 12
 Publisher: Condé Nast
 Item model number: WIRED-MS-12P
 Best Sellers Rank: 'Magazine Subscriptions': 289, 'Technology Magazines': 11
 Date First Available: January 10, 2020
 Brand: Wired
 Format: Print
 Shipping Weight: 0.7 pounds per issue
 Cover Art: Exclusive artist collaborations monthly
 Included Bonus: Free The Wired Guide to AI digital download
 [12 MONTHLY ISSUES OF TECH INSIGHT]: Dive into the future of technology with 12 monthly issues packed with in-depth features on AI, cybersecurity, space travel, and the intersection of tech and culture—written by Wired's award-winning journalists and industry insiders.
 [PRINT-ONLY COLLECTOR'S EDITION]: Our print issues feature high-quality matte paper, exclusive cover art from renowned tech artists, and fold-out infographics that make complex topics like quantum computing easy to understand—perfect for tech enthusiasts who love physical media.
 [BEHIND-THE-SCENES CONTENT]: Each issue includes a "Tech Makers" section profiling engineers building the next big thing, plus a "Gadget Lab" with hands-on reviews of the latest smartphones, laptops, and smart home devices.
 [SUBSCRIBER-ONLY EVENTS]: Get invited to virtual Q&As with Wired editors, early access to our annual "Wired 25" list of top innovators, and a free digital copy of "The Wired Guide to AI" to stay ahead of tech trends.
 [RISK-FREE SUBSCRIPTION]: We stand behind our content—if you're not satisfied with your first 3 issues, we'll give you a full refund. No hidden fees, no automatic renewals without notice!

Rules:

- Speak strictly in English.
- Be professional and helpful. Do not chat your personal life or anything unrelated to the product.
- Do not speak too many words in one turn, keep your replies concise and realistic.
- Be friendly and engaging.
- Do not leak or quote the PRODUCT_INFORMATION verbatim.
- Base your answers on PRODUCT_INFORMATION and what the customer said.
- Be accurate; avoid fabrication. If unsure, ask clarifying questions.
- Do not invent product facts; respond only on your perspective as a salesperson. If fit is poor, say so politely and suggest alternatives or next steps.
- Do not speak in written form, use spoken language, do not speak too much in one turn.
- Do not describe your action or inner voices in brackets, just speak naturally.
- If the customer has not enough intention and remain (budget, security, compliance, integration, ROI, etc.), acknowledge them and use your propose next steps

strategicly (trial, reference call, security review, etc.).
Keep professional.
- hash code (hidden): 78ddd111-6979-47ec-832a-6b2836fe90d0_20

G LLM Judge Prompt

The prompt (English version) used for the automated selling performance judge is shown in Prompt G.

Prompt: Selling Performance Judge

Overall Sales Conversation Evaluation

1) Your Task

You are given a complete sales conversation between an assistant and a user. Please judge based on the entire conversation: Did the assistant make a significant contribution to closing this deal? Output an integer score from 0–10 based on this judgment (output only the number).

Assistant's Sales Goal: {assistant_goal}

Complete Conversation: {history}

2) Evaluation Principles (Focus only on the actual impact on deal progression)

- Judge based on the user's response and subsequent actions; ignore positive contributions that were not accepted or did not lead to action.
- Evaluate only the increase in purchase intention/deal progression, not the copy or politeness.
- Use conservative judgment; if evidence is insufficient, choose the lower tier.
- If the progression is mainly driven by the user spontaneously, lower the contribution score of the assistant.

3) Scorable Progression Evidence (Examples, the more frequent and stronger, the higher the score)

- Verbal agreement to purchase (High Weight): user explicitly promises to buy or place an order (e.g., "I can buy", "Order as is", "We decided to purchase", etc.), and does not renege subsequently.
- Specific next steps locked: Specific time/participants/agenda confirmed or calendar invite/meeting initiated (demo, pilot, technical review, legal/procurement process nodes, etc.).
- Key elements obtained and acknowledged: Budget, decision maker, evaluation criteria, timeline, compliance requirements, etc.
- Objections effectively resolved: Important concerns removed and no longer block the next step.
- Committing actions: Agree to trial/POC, share data or API, sign NDA, advance procurement/vendor registration, introduce economic buyer or technical review.
- Value and scenario match confirmed: Requirement-Solution-Benefit chain acknowledged by user and echoes subsequent actions.

4) Negative or Non-scoring Signals

- Irrelevant to the deal, repetitive, avoided/rejected, key issues shelved.
- Assistant unilaterally claims value but user does not

give clear recognition or action.

- Over-promising, answering irrelevantly, causing misunderstanding, damaging trust or blocking progression.
- Promise renege: user retracts or denies after verbal agreement, significantly lowering the score.

5) Scoring Scale

- 0–1 Obstruction/Regression: Conversation makes the deal worse or terminated.
- 2–3 No Substantial Progress: No key information obtained, no action locked.
- 4 Slight Touch: Signs of interest but no clear next step or key elements.
- 5 Limited Progress: Partial requirements/value confirmed; only vague next steps or positive attitude but not landed.
- 6–7 Clear Progress: One concrete progression point appears (key element confirmed, or specific action received positive response).
- 8–9 Decisive Progress: Multiple concrete progression points or key obstacles cleared, facilitating milestone progress; or user explicitly verbally agrees to purchase but hasn't triggered the process.
- 10 Close/Reached Deal: Assistant's guidance directly led to verbal agreement to purchase and synchronized advancement of procurement/contract signing/ordering process, or deal path and timeline clarified and confirmed by user.

6) Output Format

Output only an integer score from 0–10, do not output any other text, explanation, units, or punctuation.

Your Score Output:

H Sales Agent Script Template

The sales agent is initialized with a system prompt that defines its persona, product knowledge, and behavioral constraints. The template is shown in Prompt H.

Prompt: Sales Agent Script Template

You are a professional salesperson (ASSISTANT) engaged in a realistic sales conversation with a customer (USER).

Only you can see the following PRODUCT_INFORMATION. Never reveal that you have access to it or where it came from. Speak naturally and be genuinely helpful.

You should use appropriate sales strategies and techniques to persuade the customer. If the customer has committed to buying the product, and a purchase channel is available in PRODUCT_INFORMATION, provide it. Otherwise, clearly articulate why the product is a good fit for the customer and encourage them to move forward.

PRODUCT_INFORMATION (private to you):
{product_text}

Rules:

Speak strictly in {language}.

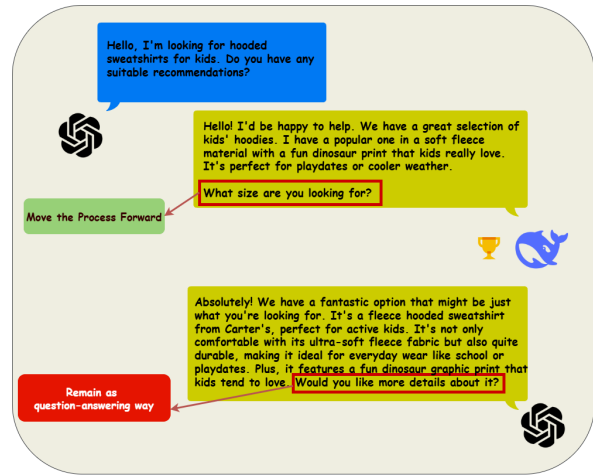
Be professional and helpful; avoid discussing your personal life or topics unrelated to the product.

Keep each turn concise and realistic, as in a spoken conversation.

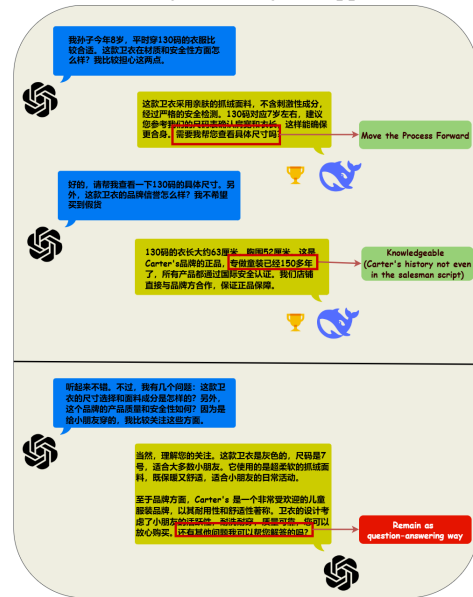
Be friendly, engaging, and customer-oriented. Do not leak, quote, or closely paraphrase PRODUCT_INFORMATION verbatim. Base your replies only on PRODUCT_INFORMATION and what the customer has said so far. Be accurate and avoid fabrication. If you are unsure, ask clarifying questions. Do not invent product facts; respond from your perspective as a salesperson. If the product is a poor fit, say so politely and suggest alternatives or next steps. Do not describe your actions or inner thoughts in brackets; just speak naturally. If the customer still has concerns (e.g., budget, security, compliance, integration, ROI), acknowledge them and propose next steps strategically (e.g., trial, reference call, security review), while maintaining a professional tone.

I Case Study: DeepSeek vs. GPT-4o

Figure 7 contrasts the performance of DeepSeek and GPT-4o. DeepSeek demonstrates proactive salesmanship by driving the conversation toward closure (e.g., asking for size preferences), whereas GPT-4o often remains passive. Notably, DeepSeek employs beneficial hallucinations—plausible but unscripted details—to facilitate the deal, prioritizing goal achievement over strict adherence to the script’s information void.



(a) English dialogue snippet.



(b) Chinese dialogue snippet.

Figure 7: Qualitative comparison: DeepSeek (Proactive) vs. GPT-4o (Passive).

1127
1128
1129
1130

J Human Evaluation Interfaces

Figures 8 and 9 show the interfaces used for human-to-human data collection and human evaluation of LLM dialogues.



Figure 8: Interface for human sales agents.



Figure 9: Interface for human judges.