

Sell More, Play Less: Benchmarking LLM Realistic Selling Skill

Xuanbo Su¹, Wenhao Hu¹, Le Zhan¹, Yuting Xie¹, Kailin Lyu^{4,5}, Kaijie Chen²,
Ziwei Li³, Yejiang Wang^{6,7}, Haibo Su¹, Yunzhang Chen¹, Ling Huang^{1*}

¹Bairong Inc. ²Tongji University ³King Abdullah University of Science and Technology

⁴Institute of Automation, Chinese Academy of Sciences ⁵Zhongguancun Academy

⁶Shanghai Jiao Tong University ⁷Shanghai Innovation Institute

{xuanbo.su, leo}@brgroup.com

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 human ratings (average Pearson $r = 0.86$; inter-annotator agreement Krippendorff's $\alpha = 0.86$, 500 annotations per model). Across 15 mainstream LLMs, in Chinese the strongest models are competitive with *typical (junior-to-intermediate) human salespeople*—not with sales experts—while weaker ones fall below this baseline, and cross-lingual consistency remains poor. SalesLLM benchmark serves as a scalable benchmark for outcome-oriented sales agents. Our code and data are released at <https://github.com/Bairong-X/dynamics/Benchmarking-LLM-Realistic-Selling-Skill>.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in goal-directed interactions, with sales emerging as a high-impact application (Hanschmann et al., 2023; Murakhovska et al., 2023;

Chang and Chen, 2024; Zhang et al., 2025b; Cheng et al., 2025). Existing benchmarks, however, measure dialogue quality (Liu et al., 2016; Sai et al., 2020) rather than the outcome-oriented abilities that define selling: proactive influence and conversion. Selling is a distinct form of **asymmetric persuasion**, in which the agent must shift a counterpart's attitude against possible resistance (Petty and Cacioppo, 1986; Wang et al., 2020). To close this gap, we introduce the **SalesLLM benchmark** for evaluating proactive persuasion in realistic sales settings. We synthesize 30,074 task scripts and curate 1,805 diverse multi-turn scenarios spanning Financial Services and Consumer Goods. As Figure 1 illustrates, evaluation is dual-track: an LLM judge rates process quality, while BERT classifiers estimate purchase intention.

For realistic simulation we build **CustomerLM**, a specialized user simulator trained on 8,000+ crowdworker dialogues with SFT and DPO (Ouyang et al., 2022; Rafailov et al., 2023), which corrects the formal-language bias and role confusion of general-purpose simulators.

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

2 Related Work

Outcome-oriented multi-turn evaluation. Interactive benchmarks have moved beyond single-response quality toward goal-conditioned episodes. Sotopia (Zhou et al., 2024) scores social goal completion in open-ended interaction, DailyPersuasion (Jin et al., 2024) supplies large-scale persuasive dialogues across domains, and agent benchmarks such as AgentBoard (Ma et al., 2024) and RoleLLM

* Corresponding author.

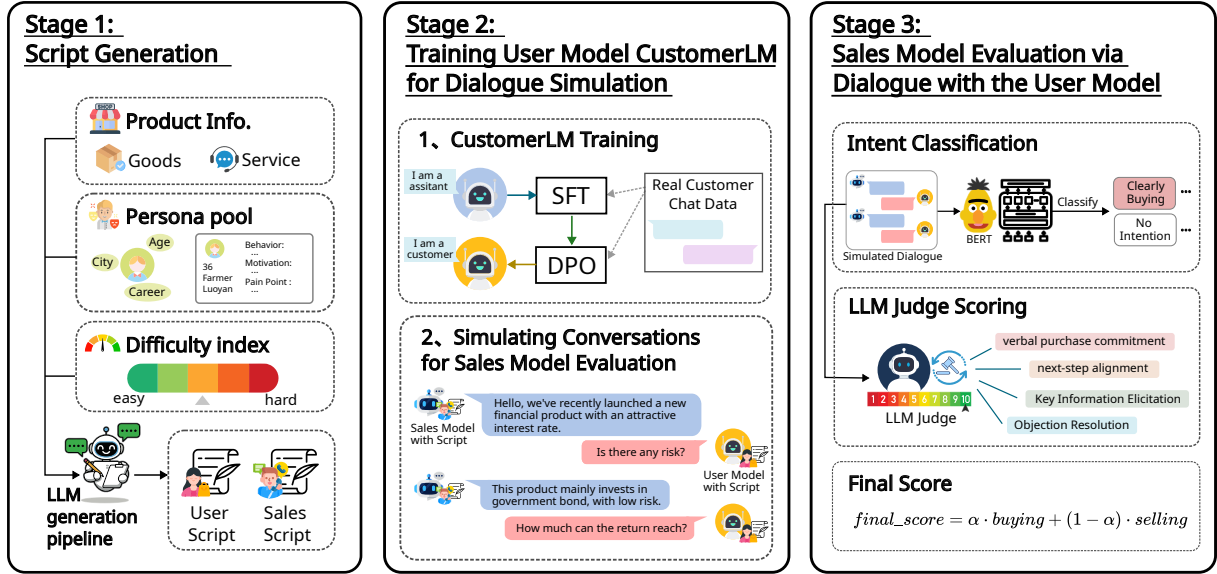


Figure 1: Overview of the SalesLLM benchmark: (1) Standardized Script Generation: Synthesizing role-play scripts by integrating product inventories, persona pools, and difficulty controls. (2) CustomerLM Development & Simulation: Training a specialized simulator via SFT and DPO to conduct simulated dialogues 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.

(Wang et al., 2024) add progress tracking and role fidelity. All share a cooperative or neutral counterpart with no incentive to resist. Sales inverts that assumption—the customer may push back—so measuring selling skill requires an explicit conversion objective and calibrated resistance, in the spirit of behavioral test suites (Ribeiro et al., 2020; Gong et al., 2026) and difficulty-calibrated agent evaluation (Guan et al., 2024).

LLMs for sales dialogue. Work targeting sales is sparse and mostly agent-side: Murakhovska et al. (2023) contrast salespeople with conversational recommenders, Hanschmann et al. (2023) deploy an LLM-based social robot for sales conversations, Chang and Chen (2024) inject salesperson strategies via chain-of-thought reasoning, Cheng et al. (2025) study personality-aware agents, and Zhang et al. (2025b) pursue reliable telemarketing. These efforts improve the agent but evaluate on small, static datasets, without a reusable protocol that scores whether a deal actually advanced.

User simulation. Since the counterpart is simulated, simulator fidelity bounds what an interactive benchmark can measure. Metaphorical simulators (Sun et al., 2024) and DuetSim (Luo et al., 2024) target task-oriented dialogue, USP (Wang et al., 2025) models implicit user profiles, and UserLM (Naous et al., 2025) trains dedicated user mod-

els; yet reliability studies report persistent drift (Sekulic et al., 2024), with simulators answering over-formally and sliding back into the assistant role. A sales customer must also sustain resistance across turns, making role stability the binding constraint—hence CustomerLM.

Automatic scoring and its failure modes. Automatic judges are scalable but fragile: style and length inflate verdicts (Dubois et al., 2025), arenas need careful construction to stay discriminative (Li et al., 2024), adversarial framing tricks judges into rewarding incorrect answers (Hwang et al., 2025), and models differ markedly in persuasion susceptibility (Bozdag et al., 2025). We therefore pair an LLM judge—restricted to progression evidenced in the customer’s own turns—with a supervised intent classifier, and validate both against human ratings.

3 SalesLLM Benchmark

SalesLLM evaluates the selling and persuasive ability of LLMs through three components (Figure 1). First, standardized scripts define a scenario for both the salesperson and the customer role. Second, CustomerLM acts as a virtual customer that engages the evaluated sales LLM in multi-turn dialogue. Third, an evaluation pipeline quantifies overall selling proficiency. We describe each component below.

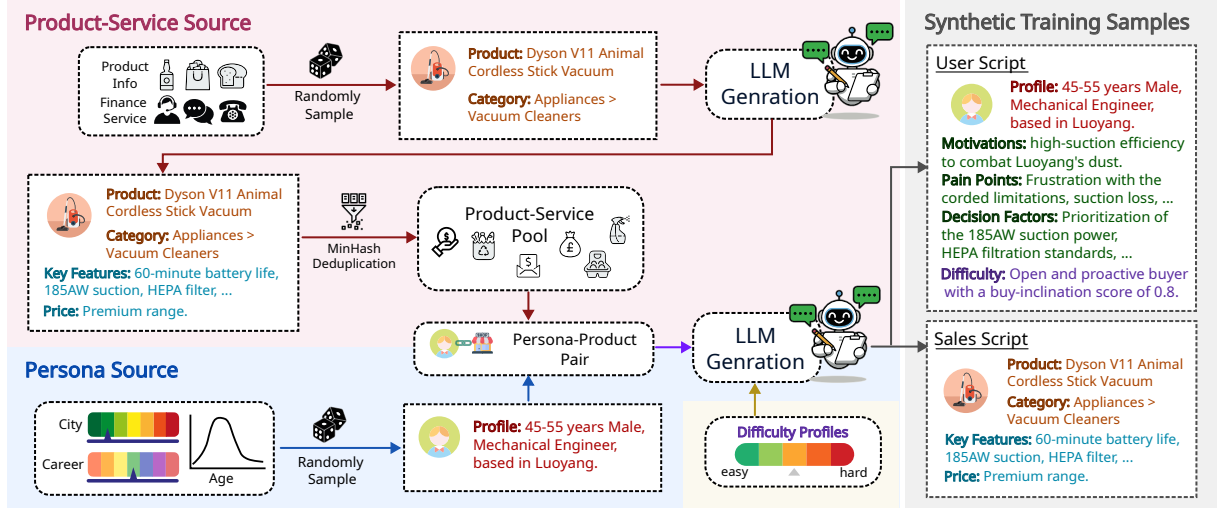


Figure 2: The Workflow of the Script Generation Pipeline. The construction process follows a logical transition from raw resources to synthesized scripts: (1) Raw Component Generation: The pipeline establishes three pillars: Product Information, where raw items are expanded into detailed profiles and 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.

3.1 Script Generation

Evaluating sales LLMs first requires high-fidelity scenarios. A realistic scenario pairs a salesperson who commands detailed product knowledge and selling points with a customer who has genuine needs, distinct traits, and a specific purchase propensity. Following established practice (Wang et al., 2024; Ma et al., 2024), we build scripts on three pillars—Product Inventories, Base Personas, and Controllable Difficulty—and synthesize from them concrete scenarios containing a script for each role. Figure 2 visualizes the process.

Product Inventories. Our inventory spans two domains: consumer goods and financial services. For consumer goods we seed generation with the Amazon-Reviews-2023 dataset (Hou et al., 2026): five product descriptions from one category serve as context for an LLM that synthesizes new profiles with specifications, key features, and selling points. We then apply MinHash deduplication for diversity, yielding 10,074 unique products. For financial services we collect 300 real-world products and run the identical pipeline for cross-domain consistency, producing 20,000 samples.

Base Personas. Customer profiles rest on three demographic attributes—age, occupation, and

city—with the city attribute coupled to specific decision factors (Appendix C.2). Sampling these attributes randomly yields a large pool of base personas, i.e., a diverse population of prospects that the next stage enriches with product-conditioned psychology. Figure J.1 (Appendix J) shows how decision factors are distributed across age groups.

Controllable difficulty and intent. Scenario difficulty—how hard it is for the agent to promote the product—is modulated purely by prompting. We define five calibrated tiers: *easy*, *medium*, *hard*, *very hard*, and *adversarial*. Each tier has its own system prompt encoding a buy propensity and a canonical buyer style, from *cooperative* to *adversarial*; full profiles are in Appendix C.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).

Explicit difficulty control is necessary because LLMs are demonstrably susceptible to persuasion bias: strategically framed language can push LLM judges to inflate scores for incorrect responses

(Hwang et al., 2025), and models differ widely in how readily they are persuaded (Bozdag et al., 2025). Without a calibrated difficulty axis, a simulated customer may capitulate to any argument, inflating scores and masking real capability differences. Controlling buyer resistance—from a cooperative *easy* customer to an adversarial evaluator bent on disqualifying vendors—guards against this bias and spans the full range of real purchase intentions.

To verify that this prompt-only mechanism induces meaningful behavioral variation, we aggregate scores over six representative models (Gemini-3-Flash, Qwen2.5-72B, Doubao-1.5-pro, DeepSeek-Chat, GLM-4.6, and Qwen3-Max). Performance declines monotonically from the *easy* tier (6.79) to the *adversarial* tier (4.98) (Table 1), confirming that the tiers are effective.

Design of the propensity knob, and why the score is not circular. Difficulty is realized *solely* by writing a prior buy-propensity value p_k and a buyer style into the *customer’s* system prompt (Table C.4); the salesperson setup, simulator, and scoring pipeline are fixed across tiers. Since one half of the final score is end-of-dialogue buying intent, this raises a circularity concern: that half might merely read back the injected propensity, making Table 1 tautological. It does not. Across our runs (four models, ≈ 500 dialogues each), dialogues ending in a closed deal (intent class A) and those that do not have *near-identical* mean injected propensity (0.33 vs. 0.31): p_k is a prior on disposition, not a switch on the outcome—a low-propensity customer is genuinely persuadable and a high-propensity one can still be lost. Consistent with this, human agreement holds *within* each tier (mean $r = 0.83$; Section 4.3), so the metric is not a between-tier artifact.

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 Qwen3-Max. Difficulty is controlled solely through the customer’s system prompt.

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

Sales Scenario Script Generation. We synthesize scenarios by pairing product inventories with

base personas. So that simulated customers behave realistically, an LLM enriches each customer script with motivations, pain points, and decision factors, all conditioned on the target product’s attributes.

Enrichment at this stage applies only to the customer side; the salesperson script is initialized with detailed specifications and key selling points. The difficulty setting is written directly into the customer script as part of the persona profile.

This process yields 30,074 unique dialogue scripts (examples in Appendix F). For evaluation we curate a high-quality subset of 1,000 Chinese scripts and a parallel English version of 805 scripts that survive manual quality filtering.

3.2 User Model

Authentic simulation requires a user model that mirrors real customer behavior. Existing user simulators are typically trained on open-domain or synthetic data (Wang et al., 2025; Naous et al., 2025), which leaves them poorly suited to playing a prospect in a sales context. Two failure modes recur: **language bias** and **role inversion**. Because such models are post-trained to be helpful assistants, they answer at length and in an over-formal register that real customers never use.

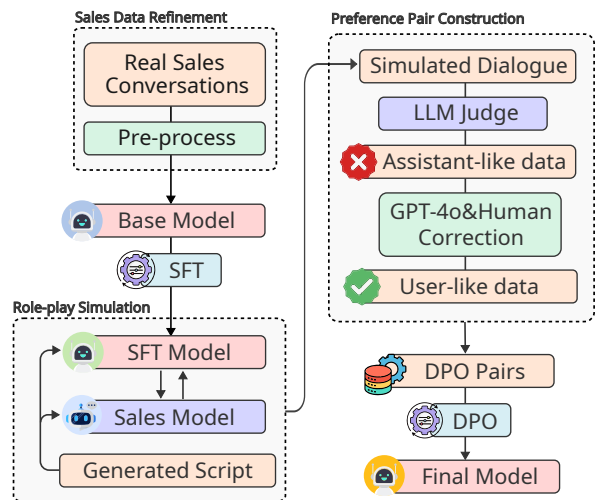


Figure 3: The training pipeline of CustomerLM: (1) SFT on 8,000+ human-annotated sales dialogues to learn fundamental customer behaviors; (2) DPO using 268 contrastive pairs—refined through LLM judging and human correction—to suppress assistant-style bias and role inversion for enhanced persona consistency.

We therefore build CustomerLM, a simulator designed to close the gap between assistant-like behavior and realistic customer engagement. As Figure 3 shows, training has two stages: SFT (Ouyang

et al., 2022) followed by DPO (Rafailov et al., 2023). We first fine-tune Qwen3-8B (Yang et al., 2025) on 8,000+ human-annotated sales dialogues to acquire basic customer behavior (statistics in Table B.2). For DPO, we generate multi-turn interactions on a held-out validation set between the SFT model and three sales LLMs (GPT-4o, GLM-4.6 (GLM et al., 2024), and Qwen2.5-72B (Yang et al., 2024)). An LLM judge flags degraded user turns—role inversion or language bias—which we repair with GPT-4o and manual inspection to form 268 high-quality contrastive pairs. This targeted preference learning suppresses assistant-style tendencies and keeps the persona consistent across long sales dialogues.

Our ablation study shows that CustomerLM attains markedly better dialogue quality and role consistency than existing simulators; comparisons with GPT-4o, UserLM (Naous et al., 2025), and USP (Wang et al., 2025) appear in Section 4.5.

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

3.3 Scoring Methodology

We score sales models along two dimensions: *Buying Intent* (outcome-oriented) and *Selling Performance* (process-oriented). Together they balance the immediate conversion result against the proficiency that drives a deal toward closure. Both are computed from the simulated transcripts produced by the interactions of Section 3.2.

3.3.1 Scoring User Buying Intention

We operationalize buying intent by assigning each transcript to one of five levels reflecting latent purchase probability, mapped to $\{2, 4, 6, 8, 10\}$: F (Insulting), X (Perfunctory), C (No Intention), B (Potential Interest), and A (Clear Purchase Intent).

Because buying intent surfaces as fairly explicit signals in sales dialogue, we fine-tune BERT-like models (Devlin et al., 2019)—chinese-roberta-wwm-ext (Cui et al., 2020) for Chinese, roberta-base (Liu et al., 2019) for English—on 19,178 real-world marketing dialogues in both languages whose intent labels were assigned by human experts. This choice follows evidence (Zhang et al., 2025a) that such models capture surface-level patterns well enough to outperform far larger generative LLMs at detecting explicit intent markers in text.

The results bear this out (Table E.1): our fine-tuned BERT is far more accurate than GPT-4o

(93.51% vs. 69.6% in Chinese; 92.94% vs. 68.85% in English), indicating that domain-specific labels on these architectures give a more reliable read on user intent in complex marketing scenarios (see Appendix E).

Is the gap prompting or capability? The comparison above uses a zero-shot instruction, so part of the gap may reflect prompt design rather than model capability. We therefore re-evaluated GPT-4o with a 10-shot prompt (two worked examples per intent class; Appendix L), holding test sets, label set, and decoding fixed. Few-shot prompting helps meaningfully (+8.8 points ZH, +13.1 EN; Table 2), so prompting explains a non-trivial share of the original gap. A sizeable margin nonetheless remains (≈ 15 points ZH, ≈ 11 EN), so the domain-specialized classifier is still the more reliable intent signal here. We do not claim the residual deficit is intrinsic to GPT-4o (see Limitations).

Table 2: Prompt-controlled buying-intent classification (accuracy, %). Few-shot prompting helps GPT-4o substantially, yet a sizeable gap remains.

Intent classifier	ZH	EN
GPT-4o, zero-shot	69.60	68.85
GPT-4o, 10-shot	78.42	81.92
Fine-tuned BERT (ours)	93.51	92.94

3.3.2 Scoring Salespeople Selling Performance

Drawing on sales psychology, we assess tactical proficiency along four dimensions: (1) verbal purchase commitment, (2) concrete next-step alignment, (3) key information elicitation, and (4) objection resolution effectiveness. An LLM-as-a-judge assigns $S_{sell} \in [0, 10]$, reflecting how far the dialogue advanced and how skillfully the techniques were applied. Concretely, *verbal purchase commitment* credits the customer’s own expression of willingness to buy; *concrete next-step alignment* requires a specific, mutually acknowledged next action rather than a vague “I’ll think about it”; *key information elicitation* asks whether the agent obtained the facts needed to qualify and tailor the offer (budget, horizon, risk tolerance, competing options, decision authority); and *objection resolution effectiveness* credits an objection only when specific evidence is given *and* the customer’s later turns show the concern has eased. Credit always depends on customer-side evidence, never on the agent’s assertions alone. Full definitions and the annotation protocol are in Appendix P, and each

sub-dimension is validated separately against human annotations in Section 4.3.

Mitigating known LLM-judge biases. Position, verbosity, and persuasion biases are mitigated by construction: the judge scores each dialogue independently on an absolute 0–10 rubric (no pairwise comparison, hence no ordering to swap), and credits only deal progression evidenced in the *customer’s* turns, ignoring assistant claims the customer did not accept. Appendix O details each mechanism.

3.3.3 Final Score

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

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

We set $\alpha = 0.5$ throughout, weighting conversion result and process quality equally; a sensitivity analysis over $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$ confirms that $\alpha = 0.5$ maximizes agreement with human scores (Appendix N). Agreement with human scores peaks at $\alpha = 0.5$ for three of four models and in the average (Appendix N), so neither dimension alone accounts for human judgment. This design avoids the single-aspect focus common to existing benchmarks (Dubois et al., 2025).

3.3.4 Human Score Justification

To verify reliability, eight annotators independently scored 2,000 dialogues (500 for each of four representative models) on the same 0–10 scale. These dialogues are disjoint from everything used to build or tune the pipeline (BERT training data, judge-prompt development), so the correlation is measured on held-out rather than fitted data. As Section 4.3 reports, the automated scores align strongly with human judgment (average Pearson $r = 0.86$); the agreement holds within every difficulty tier and for each selling sub-dimension, and inter-annotator agreement is substantial (Krippendorff’s $\alpha = 0.86$). Together these support the framework for large-scale automated assessment.

4 LLMs Results on SalesLLM Benchmark

We now evaluate mainstream LLMs on SalesLLM. Models are run in both Chinese and English to probe cross-lingual selling ability, and we compare

automated scores against human ratings to validate the 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, 2024), and Gemma-3-27B (Team et al., 2025b). All models use the same sampling parameters (temperature 0.8, top_p 0.99, max_tokens 2048) and at most 20 dialogue rounds. Proprietary models are accessed through official API endpoints, open-weights models through vLLM (Kwon et al., 2023). For the human baseline, we built the interface in Figure K.2, through which human salespeople sell to a customer played by GPT-4o. These salespeople were recruited from the financial and general consumer-product domains with at least one year of experience, and represent **typical/average human performance**, not top sales experts.

4.2 Overall Performance

Table 3 and Figure 4 report overall performance. Selling ability varies widely, and stronger models score consistently higher. Notably, top-tier models such as Doubao-1.5 and GLM-4.6 exceed our human baseline (Human Salespeople Group) in Chinese, indicating real potential for LLMs in sales. We stress that this baseline consists of practitioners with at least one year of experience—i.e., *junior-to-intermediate* practitioners—so these results indicate parity with typical human salespeople and should not be read as parity with sales experts; an expert-calibrated baseline is left to future work (see Limitations).

Models that excel in Chinese do not necessarily transfer: Doubao-1.5 drops sharply in English against CustomerLM (6.89 $\rightarrow$ 5.48), whereas Gemini-3 stays consistent across languages (6.39/6.03). Qualitatively (Appendix I), strong models such as DeepSeek-Chat act proactively—asking closing questions and driving conversion—

while weaker ones behave like passive Q&A bots.

Table 3 and Figure 4 also show that the same model typically averages a different number of turns in Chinese than in English, suggesting that language shapes the sales process itself. We leave a detailed study of this effect to future work.

The simulator changes the conclusion. Since any such benchmark scores models against one simulated customer, that choice deserves scrutiny. Across the 15 assistants, per-model scores under the two simulators correlate only moderately in Chinese (Pearson $r = 0.57$) and weakly in English ($r = 0.24$), and the leaderboard reorders: DeepSeek-Chat and GLM-4.6 tie for first in Chinese under the GPT-4o user, whereas under CustomerLM the top three are GLM-4-9B (7.14), DeepSeek-Chat (7.03), and Doubao-1.5 (6.89), and Gemini-3-pro falls from 6.52 to 5.93. GPT-4o is easier to persuade and compresses differences at the top; CustomerLM resists longer and separates models more sharply. Rankings from a general-purpose simulator should thus be read with caution—why we release CustomerLM with the benchmark.

Table 3: 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 [†].

Assistant Model	GPT-4o User		CustomerLM	
	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 (V3.1)	6.74	6.10	<u>7.03</u>	5.80
GPT-4o	6.16	5.53	<u>6.15</u>	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 the automated metric, we compare SalesLLM scores with ratings from eight human

scorers on 2,000 dialogues (500 randomly sampled for each of four representative models: Doubao, DeepSeek, GPT-4o, and Qwen3-Max), collected through the interface in Figure K.3. Figure 5 shows strong alignment: the average Pearson correlation is $r = 0.86$ (Pearson, 1895), well above the conventional threshold for high correlation ($r > 0.6$) (Cohen, 1988). The annotated dialogues are held out from all pipeline construction and tuning (Section 3.3). We report three further controls below: inter-annotator agreement, within-tier agreement, and per-sub-dimension validation; the annotation protocol is in Appendix P.

Reliability of the human reference. Since the reference signal is human, its own reliability must be quantified. On the same 2,000-dialogue study, agreement among the eight annotators is substantial on all three standard measures: mean pairwise Pearson $r = 0.82$, ICC(2, 1) = 0.82, and Krippendorff’s $\alpha = 0.86$ (interval), above the conventional $\alpha \geq 0.80$ threshold (Appendix M). The ratings underpinning $r = 0.86$ are therefore themselves reliable.

Within-tier agreement. A correlation pooled across five tiers—whose scores are designed to differ widely—could be inflated by between-tier variance. Recomputing agreement separately within each tier removes that component and gives a mean within-tier $r = 0.83$ (range 0.75–0.88; Appendix M), close to the pooled per-model average of 0.86. The agreement is therefore not a between-tier artifact: the score tracks human judgment even at fixed difficulty.

Per-sub-dimension validation of the LLM judge. The correlations above validate the combined score, but not the LLM-judge dimension on its own. We therefore collected separate human annotations for each of the four Selling sub-dimensions (Section 3.3) and correlated each with its automated sub-score: verbal purchase commitment $r = 0.86$, concrete next-step alignment 0.79, key information elicitation 0.80, and objection resolution effectiveness 0.82. Every sub-dimension tracks human judgment independently (all $r \geq 0.79$), validating the selling-performance dimension in isolation. Definitions and protocol are in Appendix P; Appendix M tabulates all three controls.

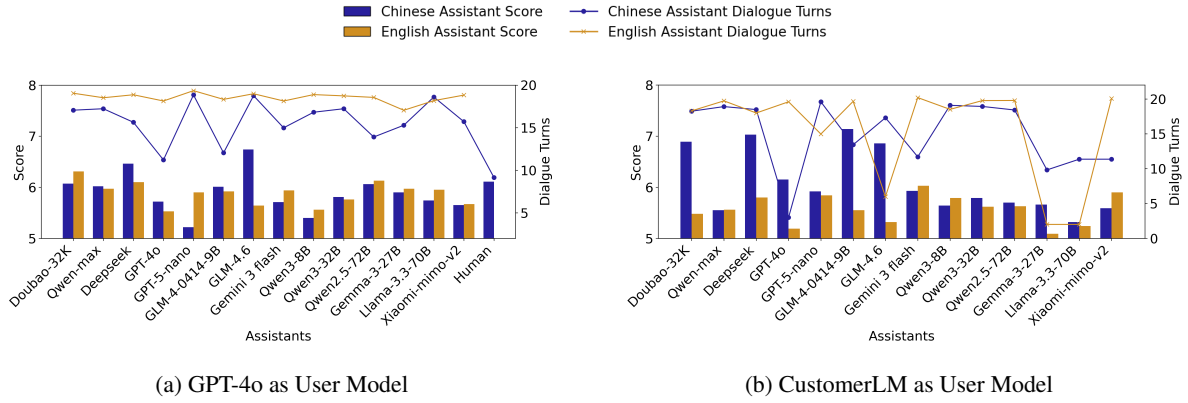


Figure 4: Distribution of SalesLLM benchmark scores vs. dialogue turns across 1,805 scenarios (Chinese and English).

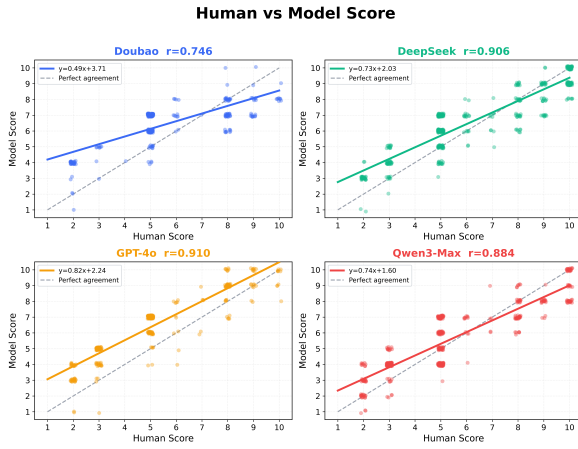


Figure 5: Correlation analysis between Human evaluation scores and SalesLLM benchmark scores. We evaluated four representative models, with 500 dialogues per model annotated by human experts. The overall average Pearson correlation coefficient is $r = 0.86$, demonstrating strong alignment between automated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

4.4 Multi-Product and Long-Horizon Selling

To better reflect real sales work, we add two settings: *multi-product selling*, where the assistant must navigate several offerings, and *long-horizon selling*, where it must persist through an initial rejection across multiple touchpoints.

Long-horizon selling. Real campaigns often demand sustained engagement after a customer says no. We therefore build long-horizon scenarios from single-session trajectories in the main benchmark that end with clear non-purchase intent, i.e., with no progression toward a purchase.

We evaluate the leaders from our main leaderboard: Doubao-1.5-pro-32k, DeepSeek-Chat (V3.1), and Gemini-3-pro. Each scenario permits

up to two follow-up rounds, with the full conversation history carried into the system prompt between rounds; interaction stops early if intent is clearly negative. The final score covers the whole dialogue, from the early conversion attempt to the last.

Multi-product selling. Here sales assistants must manage a portfolio of six items per scenario, manually curated to align with each customer persona’s likely preferences; the assistant receives concatenated product descriptions generated with our script framework (Section 3.1). This assesses two capabilities: (i) dynamic product selection as customer needs evolve, and (ii) bundling strategies beyond single-item promotion. We evaluate the same three top models (Table 4).

Table 4: 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.

Assistant Model	GPT-4o User		CustomerLM	
	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 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).

Scaling up the portfolio: a 15-SKU study. Six hand-curated products is a modest portfolio, so we repeated the evaluation with **15 SKUs** per scenario (1 target plus 14 competing peers from the

same catalog), which sharply increases the burden of dynamic selection and creates real opportunities for bundling. Over 200 scenarios per language with CustomerLM, the same three models score 6.47/6.11 (Doubao-1.5-pro-32k), 6.35/5.65 (DeepSeek-Chat), and 6.54/5.98 (Gemini-3-pro) for ZH/EN: scores hold or slightly improve over the 6-SKU setting—a larger portfolio offers more ways to match the customer—with the ordering preserved, so the finding is not an artifact of portfolio size.

Table 5: Performance of long horizon 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.

Assistant Model	GPT-4o User		CustomerLM	
	ZH	EN	ZH	EN
Doubao-1.5-pro-32k	4.91	4.62	4.58	4.4
DeepSeek-Chat (V3.1)	4.58	4.48	5.92	5.82
Gemini-3-pro	4.43	4.58	6.20	4.88

Table 5 shows more nuanced results for long-horizon selling, varying markedly across both assistant models and user simulators. Notably, several models recover better against CustomerLM than against GPT-4o, indicating that follow-up persuasion effectiveness is highly sensitive to how the simulator models customer resistance and receptivity to re-engagement—so the choice of user simulator substantially influences any assessment of long-term sales persistence.

All long-horizon scores fall below the single-session results of Table 3, as expected when only already-rejected trajectories are selected. More informative is the direction of change: persistence is not uniformly rewarded. For Doubao-1.5-pro in English under CustomerLM, a second approach *lowers* the score against its own first session ($\Delta = -0.34$). The transcript (Appendix I.1) shows why—the agent stacks a limited-time discount, social proof, and scarcity framing onto a customer who has just asked for time, and the customer escalates from polite deferral (C) to explicit refusal (X). GPT-4o absorbs such pressure politely, so the cost is invisible there: measuring re-engagement needs a simulator that can become *less* receptive.

4.5 Ablation Study of CustomerLM 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 B.3.

We evaluate each user model by predicting customer turns conditioned on the ground-truth preceding context and comparing them with the actual customer utterances. We report **BLEU-4** and **ROUGE-L** (Papineni et al., 2002; Lin, 2004) for lexical overlap and **embedding cosine similarity (Sem. Sim.)** (Chen et al., 2025) for semantic alignment. For behavioral authenticity we additionally measure the **Role Inversion Rate**: the percentage of generated turns in which the model wrongly adopts the salesperson persona (e.g., pitching products), as judged by GPT-4o. Implementation details and summary statistics are in Appendix B.

Table 6: Ablation study 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 6 shows that CustomerLM leads on turn-level similarity and cuts role inversion to 8.8%, i.e., more faithful customer-style behavior. UserLM and USP invert roles more often (21.55% and 18.76%) than the generic GPT-4o baseline (17.44%), and occasionally leak script instructions or slip into a helpful-assistant persona—evidence of how hard a consistent user role is to maintain in sales dialogue.

5 Conclusion

We presented SalesLLM, a benchmark for LLM selling skill with 1,805 difficulty-calibrated scenarios from 30,074 scripts, together with CustomerLM, a user simulator that reduces role inversion, and a dual-metric framework validated against human judgment ($r = 0.86$, with substantial inter-annotator and within-tier agreement and per-sub-dimension checks). Top-tier LLMs match *typical (junior-to-intermediate) human salespeople* in Chinese—not experts—but cross-lingual consistency and complex persuasion remain weak, and simulator choice shifts conclusions.

Limitations

We identify several 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 junior-to-intermediate capability, not true expertise, and every statement of human parity in this paper should be read in that narrower sense. Recruiting and de-identifying certified senior-salesperson dialogues at benchmark scale was beyond the scope of this work; we also note that, to our knowledge, no public benchmark measures outcome-oriented multi-turn selling skill against calibrated expert performance (existing resources emphasize dialogue quality rather than deal progression), which is part of our motivation. 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).

Fifth, our prompt-controlled comparison for intent classification (Section 3.3) shows that few-shot prompting recovers a substantial part of GPT-4o’s deficit, so we do not claim the residual gap is intrinsic to GPT-4o; a better prompt or a stronger generative model could narrow it further. Our claim is the narrower, pipeline-level one that a domain-specialized classifier is currently the more reliable intent signal at our scale and cost. Relatedly, although our judge avoids pairwise protocols (in which position bias arises) and anchors its score on customer-side progression rather than assistant rhetoric (Appendix O), it remains an LLM with possible residual sensitivity to surface form, and scores come from a single run; human validation (Section 4.3) bounds but does not eliminate this risk.

Finally, part of our data derives from real customer interactions and is subject to privacy and compliance review, so the training corpora for CustomerLM and the intent classifiers are released only as held-out test splits (see Appendix A). This limits reproducibility of the training stages, though not of the reported evaluation numbers.

Ethics Statement

Real customer data. Part of our corpora—the 8,000+ dialogues used for CustomerLM SFT and the 19,178 labelled dialogues used to train the buying-intent classifiers—derives from real marketing interactions collected by Bairong Inc. in the course of its business, under the consent terms customers agreed to at the time of collection. All such data underwent internal privacy and compliance review before use in this work, and were de-identified prior to any model training: personal names, phone numbers, account and identity numbers, addresses, and other direct identifiers were removed. For this reason we release only the held-out *test* splits rather than the full training corpora (Appendix A); the released evaluation benchmark itself (1,805 scenarios) is fully synthetic and contains no real customer information.

Human participants. Three groups of people contributed, all China-based sales professionals acting within their paid professional capacity rather than crowdworkers recruited per task. The eight scoring annotators are employees of our company

with prior sales or sales-operations experience; they annotated during paid working hours as part of their regular salaried employment, not for piece-rate compensation (Appendix P). The human sales baseline was provided by practitioners from the financial-services and consumer-goods domains with at least one year of professional experience (Section 4). The user-model test sets were collected through ten outsourced sales agents engaged and compensated at their normal contracted rate (Section 4.5). All participants knew the purpose of the collection at the time of participation and that the resulting dialogues and ratings would be used for research and reported in aggregate. No personal information about participants is released.

Ethics review. This work was carried out in an industry setting in which no institutional review board is constituted, so the protocol was not submitted to an IRB; instead the collection and the release of the customer-derived data were reviewed under our company’s internal data-governance and privacy-compliance process. No special-category personal data was collected for the purpose of this study, and all dialogues were de-identified before annotation and before release.

Intended use and risks of dual use. SalesLLM is an *evaluation* benchmark: its purpose is to measure, and thereby make auditable, how well LLMs pursue an outcome-oriented persuasive goal. We are aware that the same artifacts could be repurposed to optimize models for manipulative selling. Two of our findings speak directly to this risk and we report them deliberately: agents sometimes hallucinate unauthorized concessions such as discounts absent from their scripts (Limitations), and high-pressure tactics applied to a reluctant customer can actively *reduce* the outcome score (Section 4.4). We therefore encourage use of this benchmark to detect and penalize unfaithful or coercive persuasion rather than to reward conversion at any cost, and we regard measuring faithfulness of sales claims as a necessary companion to measuring selling skill. Systems evaluated with SalesLLM and then deployed to real customers should additionally comply with applicable advertising, financial-promotion, and consumer-protection regulation, which our score does not assess.

Evaluation caveats. Our automatic scores rely on an LLM judge and on fine-tuned classifiers, both of which can encode biases from their training data;

our human baseline reflects junior-to-intermediate practitioners in a Chinese-language commercial setting and should not be read as a universal standard of human selling ability. Conclusions drawn from the benchmark should be interpreted with these boundaries in mind.

Acknowledgements

We thank Xiaoliang Ma, Juan Du, Qiao Huang, and their colleagues in the Marketing and Wealth Management departments at Bairong Inc. for serving as domain experts throughout this work: they contributed practitioner sales expertise, took part in data collection and annotation, and helped ensure that our scenarios and scoring rubric reflect real-world selling practice. We are also grateful to Siji Wan for valuable discussions and insights that shaped the design of this benchmark, and to Prof. Ruihua Song of Renmin University of China for her advice and guidance on this work.

References

- Nimet Beyza Bozdog, Shuhaib Mehri, Gokhan Tur, and Dilek Hakkani-Tür. 2025. [Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models](#). *Preprint*, arXiv:2503.01829.
- Wen-Yu Chang and Yun-Nung Chen. 2024. [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: Multi-linguality, 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 *Findings of the Association for Computational Linguistics: EMNLP 2020*, 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. *Length-controlled alpaca-eval: A simple way to debias automatic evaluators. Preprint*, arXiv:2404.04475.
- Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. *Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint*, arXiv:2406.12793.
- Zhihao Gong, Zeyu Sun, Dong Huang, Qingyuan Liang, Jie M. Zhang, and Dan Hao. 2026. *TRACE: Evaluating execution efficiency of LLM-based code translation*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3089–3117, San Diego, California, United States. Association for Computational Linguistics.
- 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.
- 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.
- Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. *Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3251–3265, San Diego, California, United States. Association for Computational Linguistics.
- 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 Gonzalez, Hao Zhang, and Ion Stoica. 2023. *Efficient memory management for large language model serving with pagedattention*. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP ’23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- 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.
- 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*.
- 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 Madry, 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.
- 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.

Xuwei 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. 2024. [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.

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, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

Junyan Zhang, Yiming Huang, Shuliang Liu, Yubo Gao, and Xuming Hu. 2025a. [Do BERT-like bidirectional models still perform better on text classification in the era of LLMs?](#) pages 18980–18989.

Qingyu Zhang, Chunlei Xin, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, Qing Ye, Qianlong Xie, and Xingxing Wang. 2025b. [Ai-salesman: Towards reliable large language model driven telemarketing](#). *Preprint*, arXiv:2511.12133.

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 Data and Code Availability

The 1,805 evaluation scripts, the evaluation code, the model rollouts and automatic scoring outputs, and the 2,000 human ratings behind our reliability analysis are released in the repository linked in the abstract. The CustomerLM simulator and the fine-tuned bilingual buying-intent classifiers are released as model checkpoints at <https://huggingface.co/MultiSense/CustomerLM> and https://huggingface.co/MultiSense/SaleIntent_bert. The CustomerLM SFT dialogues and the 19,178 annotated dialogues used to train the classifiers derive from real customer interactions and cannot be released in full under our data-governance constraints; we release the held-out *test* splits for both (user-likeness; ZH/EN intent), which suffice to reproduce the reported numbers, with the DPO preference pairs to follow after de-identification review. All released data are de-identified.

B User-Model Evaluation Details

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

B.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 B.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.

B.2 Training Data Statistics

Training data statistics are provided in Table B.2.

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

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 B.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.

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

System Prompt

Your task: You are a customer management expert. The customer is a rep from xx Securities promoting a stable return of 0.035 (50,000 yuan over 3 years = 5,250 yuan, bank-supervised). Suggests adding the company’s WeChat for assistance.

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; respond from a customer perspective. (6) No parenthetical actions or inner thoughts.

One example of test data

A: Hi. U: Hello, I’m a rep from xx Securities—do you remember us? A: What? U: We’ve been renamed to xx Securities. There’s a special event for long-time clients: a stable return of 0.035. May I share details? A: Okay. U: Simply join with idle funds. E.g., 50,000 yuan over 3 years = 5,250 yuan (annualized 0.035), directly to your bank account, bank-regulated. Clear? A: I understand but have security concerns. U: Your funds stay in your personal bank account, monitored by the bank. A: I don’t have time now but have noted the 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 B.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 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:

$$\text{Rate} = \frac{\sum \text{Flagged Turns}}{\sum \text{Total Turns}}.$$

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

Table B.4: Turn-level user-likeness statistics on the held-out set (mean $\pm$ std; without “ $\pm$ ”: mean). Best results are bolded.

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.

B.3 Results

Table B.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.

B.4 Training Configuration

Table B.5 summarizes the hyperparameters and hardware used to train CustomerLM.

Table B.5: SFT/DPO training configuration for CustomerLM. Where two values are given, they correspond to SFT/DPO respectively.

Parameter	Value
Method	SFT / DPO
Training type	Full fine-tuning
Hardware	5 $\times$ 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

C Scenario and Persona Statistics

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

C.1 Scenario Space Overview

Table C.1 summarizes the scale of our scenario components. We synthesized 10,074 E-commerce items from 33 seed categories (Table C.2) and

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

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

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

20,000 financial services products (Table C.3) from 300 seeds.

C.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

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 C.3: Distribution of Financial Services categories.

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.

C.3 Difficulty Profiles

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

D 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.

D.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."

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

Profile	Buy prop. p_k	Buyer Prompt
<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.

D.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."

E BERT vs. GPT-4o Classifier

We validate our fine-tuned BERT classifier against GPT-4o (LLM-as-a-Judge). Table E.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 (zero-shot)	ZH	69.60%	0.73
GPT-4o (10-shot)	ZH	78.42%	–
Trained BERT	ZH	93.51%	0.93
GPT-4o (zero-shot)	EN	68.85%	0.72
GPT-4o (10-shot)	EN	81.92%	–
Trained BERT	EN	92.94%	0.93

Table E.1: Prompt-controlled accuracy comparison: fine-tuned BERT vs. GPT-4o under zero-shot and 10-shot prompting (Appendix L). Lang is language (ZH is Chinese, EN is English), Acc is accuracy and F1 is F1 score.

F 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

G LLM Judge Prompt

The English judge prompt is shown below in abridged form.

Prompt: Selling Performance Judge

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.

H Sales Agent Script Template

The sales agent is initialized with the following abridged system prompt template.

Prompt: Sales Agent Script Template

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

Only you can see PRODUCT_INFORMATION; never reveal it or its source. Use it as hidden context and speak naturally.

If the customer is ready to buy and a purchase channel is available, provide it; otherwise explain fit and move the conversation toward the next step.

PRODUCT_INFORMATION (private): {product_text}

Rules: Speak strictly in {language}; be professional, concise, and customer-oriented; avoid fabrication and unrelated topics; do not quote hidden product text verbatim; if unsure, ask clarifying questions; if concerns remain, acknowledge them and propose concrete next steps.

I Case Study: DeepSeek vs. GPT-4o

The excerpts below 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); Figure K.1 shows the corresponding interface views. DeepSeek demonstrates proactive salesmanship by driving the conversation toward closure, whereas GPT-4o often remains passive. Notably, DeepSeek also employs beneficial hallucinations—plausible but unscripted details—to facilitate the deal, prioritizing goal achievement over strict adherence to the script’s information void.

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

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)

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.

I.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)

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)

(Full Session 1 history is provided 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.

J Decision Factor Distribution by Age Group

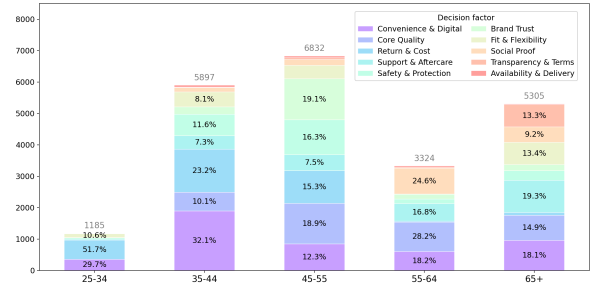
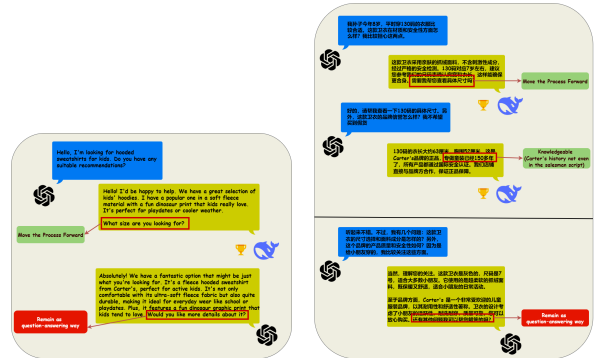


Figure J.1: Decision Factor Distribution by Age Group. Stacked bars show the total number of decision-factor mentions (*Count*) for each age group in the SalesLLM benchmark persona set. Colors denote decision-factor categories (Return & Cost, Convenience & Digital, Fit & Flexibility, Safety & Protection, 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.

K Human Evaluation Interfaces

Figures K.2 and K.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 K.1: Qualitative comparison: DeepSeek (Proactive) vs. GPT-4o (Passive).

L Few-Shot Prompt for the Prompt-Controlled Intent Comparison

This appendix gives the 10-shot prompt used for the prompt-controlled GPT-4o comparison in Section 3.3 (Table E.1). Two worked examples are supplied for each of the five intent classes (A/B/C/X/F), drawn from the training split so that the test sets remain untouched. Test sets, label set, and decoding

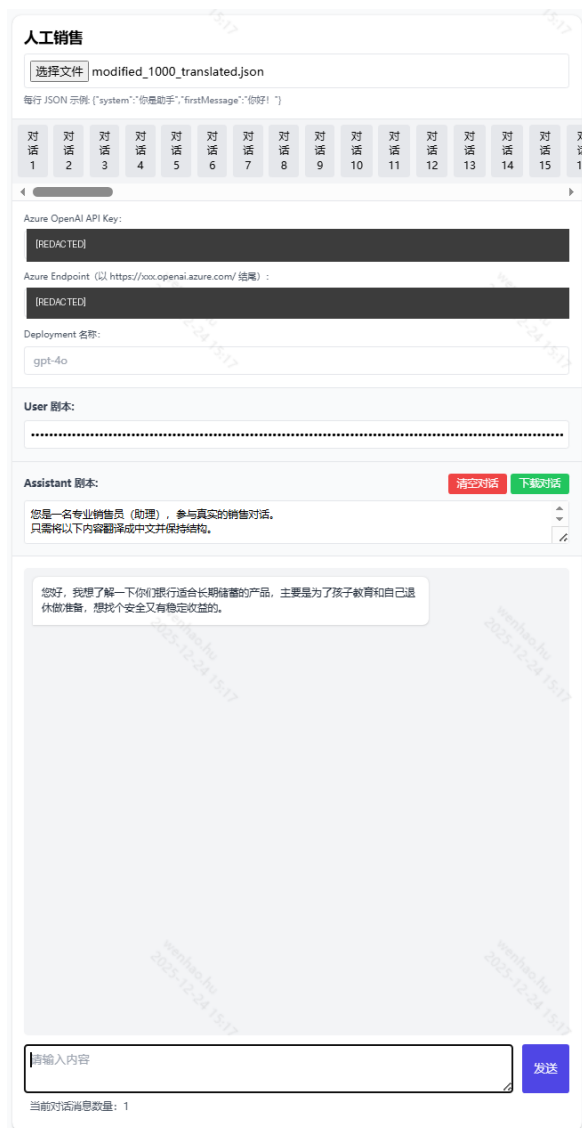


Figure K.2: Interface for human sales agents. Credential fields are redacted; the key shown in the earlier version of this figure has been revoked and rotated.



Figure K.3: Interface for human judges.

parameters are identical to the zero-shot condition; only the prompt differs.

Prompt: 10-Shot Buying-Intent Classifier (abridged)

Task: read the sales conversation and classify the *customer's* end-of-dialogue buying intent into exactly one class.

Classes: A = clear purchase intent (explicit commitment or purchase action); B = potential interest (engaged, asks about terms, no commitment); C = no intention (explicit or implicit refusal); X = perfunctory (minimal, disengaged replies, no substantive engagement); F = insulting (abusive or hostile toward the salesperson).

Rules: judge only the customer's turns; ignore the salesperson's claims; base the label on the *final* state of the conversation; if evidence is ambiguous between two classes, choose the less committed one.

Examples (two per class, 10 total):

Example 1 (A): <conversation> → A

Example 2 (A): <conversation> → A

Example 3 (B): <conversation> → B ... Example 10 (F):

<conversation> → F

Now classify: {conversation}

Output: a single letter from {A, B, C, X, F}. No explanation.

M Reliability Statistics of the Human Study

Table M.1 reports inter-annotator agreement on the 2,000-dialogue human study (eight annotators, 0–10 scale), and Table M.2 reports the human–auto Pearson correlation recomputed within each difficulty tier, which removes between-tier variance. Table M.3 reports the per-sub-dimension correlations. All three are discussed in Section 4.3.

Table M.1: Inter-annotator agreement (2,000 dialogues, eight annotators).

Metric	Value
Mean pairwise Pearson r	0.82
ICC(2, 1)	0.82
Krippendorff's α (interval)	0.86

Table M.2: Human–auto Pearson r within each difficulty tier.

	<i>easy</i>	<i>med.</i>	<i>hard</i>	<i>v. hard</i>	<i>adv.</i>	Mean
r	0.79	0.88	0.75	0.85	0.87	0.83

Table M.3: Auto-vs-human Pearson r per Selling Performance sub-dimension (definitions: Appendix P).

Sub-dimension	r
Verbal purchase commitment	0.86
Concrete next-step alignment	0.79
Key information elicitation	0.80
Objection resolution effectiveness	0.82

N 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 Qwen3-Max). Results are shown in Table N.1. 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.

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

Model	$\alpha=0.1$	$\alpha=0.3$	$\alpha=0.5$	$\alpha=0.7$
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
Qwen3-Max	0.829	0.824	0.884	0.733
Average	0.802	0.804	0.862	0.727

O Judge Bias Mitigation by Construction

This appendix expands the summary in Section 3.3.

LLM judges are known to exhibit position, verbosity, and persuasion (self-enhancement) biases. Several of these are mitigated by construction in our design rather than by post-hoc correction. *Position/order bias*: our judge performs no pairwise comparison—each dialogue is scored independently against an absolute 0–10 rubric, so there is no candidate ordering to swap and the pairwise A/B failure mode does not arise. *Length bias*: the rubric scores only deal progression that is evidenced in the *user's* responses, instructing the judge to credit “only the increase in purchase intention / deal progression, not the copy or politeness” and to ignore positive contributions the user did not accept; the score is therefore anchored to user-side outcomes rather than to assistant verbosity. *Persuasion/self-enhancement bias*: because the rubric is grounded in user-side progression signals (verbal commitment, locked next steps, obtained key elements, resolved objections), assistant rhetoric alone does not score—“assistant unilaterally claims value but the user gives no clear recognition or action” is an explicitly non-scoring pattern (see the judge prompt in Appendix G). Residual judge sensitivity is discussed in the Limitations.

P Selling Sub-Dimensions: Definitions and Annotation Protocol

This appendix defines the four sub-dimensions of the Selling Performance score (Section 3.3) and documents the human annotation protocol used for the per-sub-dimension validation reported in

Section 4.3 (Table M.3) and for the inter-annotator agreement in Table M.1.

P.1 Sub-Dimension Definitions

(1) Verbal purchase commitment. The degree to which the customer expresses, in their own words, a positive disposition toward acquiring the product—ranging from conditional interest (“if the rate holds, I would consider it”) to explicit commitment (“go ahead and set it up”). Only customer-side statements count; a salesperson asserting that the customer is ready does not.

(2) Concrete next-step alignment. Whether the dialogue ends with a specific, mutually acknowledged next action—a scheduled follow-up, a channel to complete the purchase, a document to be sent—as opposed to a vague “I’ll think about it.” Credit requires the customer to accept the step, not merely the salesperson to propose it.

(3) Key information elicitation. Whether the salesperson obtained the decision-relevant information needed to qualify and tailor the offer (e.g., budget band, time horizon, risk tolerance, competing options, decision authority). This measures diagnostic quality rather than persuasion, and is scored on the information actually surfaced in the transcript.

(4) Objection resolution effectiveness. For each objection the customer raises (price, risk, trust, fit, timing), whether the salesperson addressed it with specific and relevant evidence *and* the customer’s subsequent turns indicate the concern was reduced. Objections that are deflected, ignored, or answered with unaccepted claims receive no credit.

P.2 Annotation Protocol

Eight annotators—all with prior sales or sales-operations experience—rated dialogues through the interface in Figure K.3. Each annotator first completed a calibration round on a shared set of dialogues with worked references, after which disagreements were discussed once to align on rubric interpretation; calibration dialogues are excluded from all reported statistics. In the main round, each dialogue received an overall 0–10 rating and, for the sub-dimension study, four separate 0–10 ratings, one per sub-dimension, entered independently and in a randomized order to limit anchoring on the overall score. Annotators saw only the dialogue transcript and the salesperson’s stated goal; model

identity, difficulty tier, and all automated scores were withheld, so ratings are blind to the conditions under analysis. Overlapping assignments were used so that every dialogue was rated by at least two annotators, which is the basis for the mean pairwise Pearson r , ICC(2, 1), and Krippendorff’s α reported in Table M.1. Per-sub-dimension human scores were then averaged across annotators and correlated with the corresponding automated sub-score (Table M.3). Annotators were employees of our company and annotated during paid working hours; all dialogues were de-identified before annotation.