

Dataset

Dialogue generation

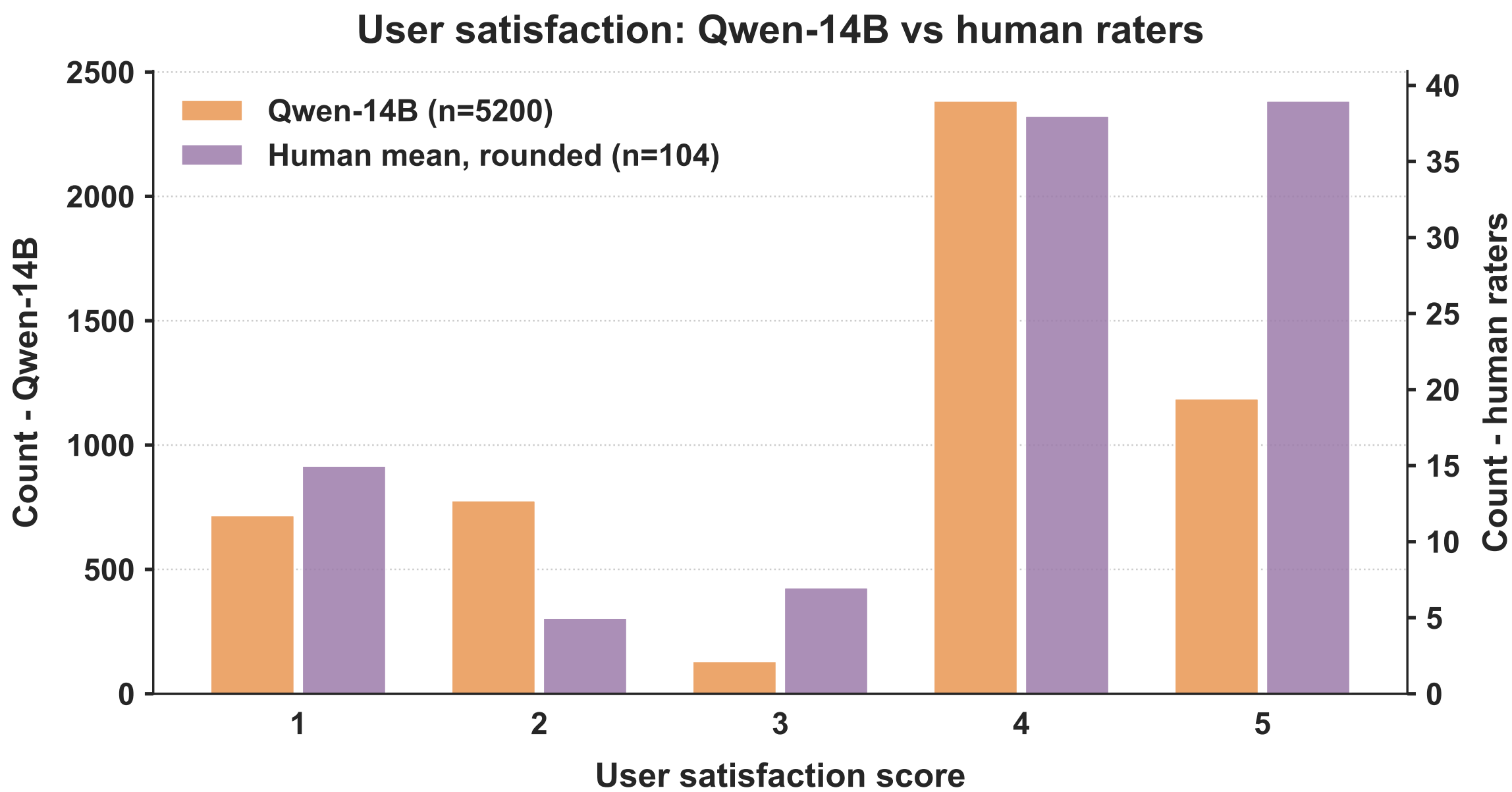
- Sales agent: GPT-4o-mini
- User simulator: Gemini-1.5-Flash
- Personas: Claude 3.5 Sonnet v2

Agent configs	13
Simulator configs	400 (100P × 1G × 4S)
Total dialogues	5200
Turns/dialogue	12.25 ± 2.78 (5–31)
Chars/turn	U: 226 ± 65, A: 707 ± 210

Table 1. Dialogue dataset. P: Personas, G: Goals, S: Scenarios, U: simulated User, A: Agent

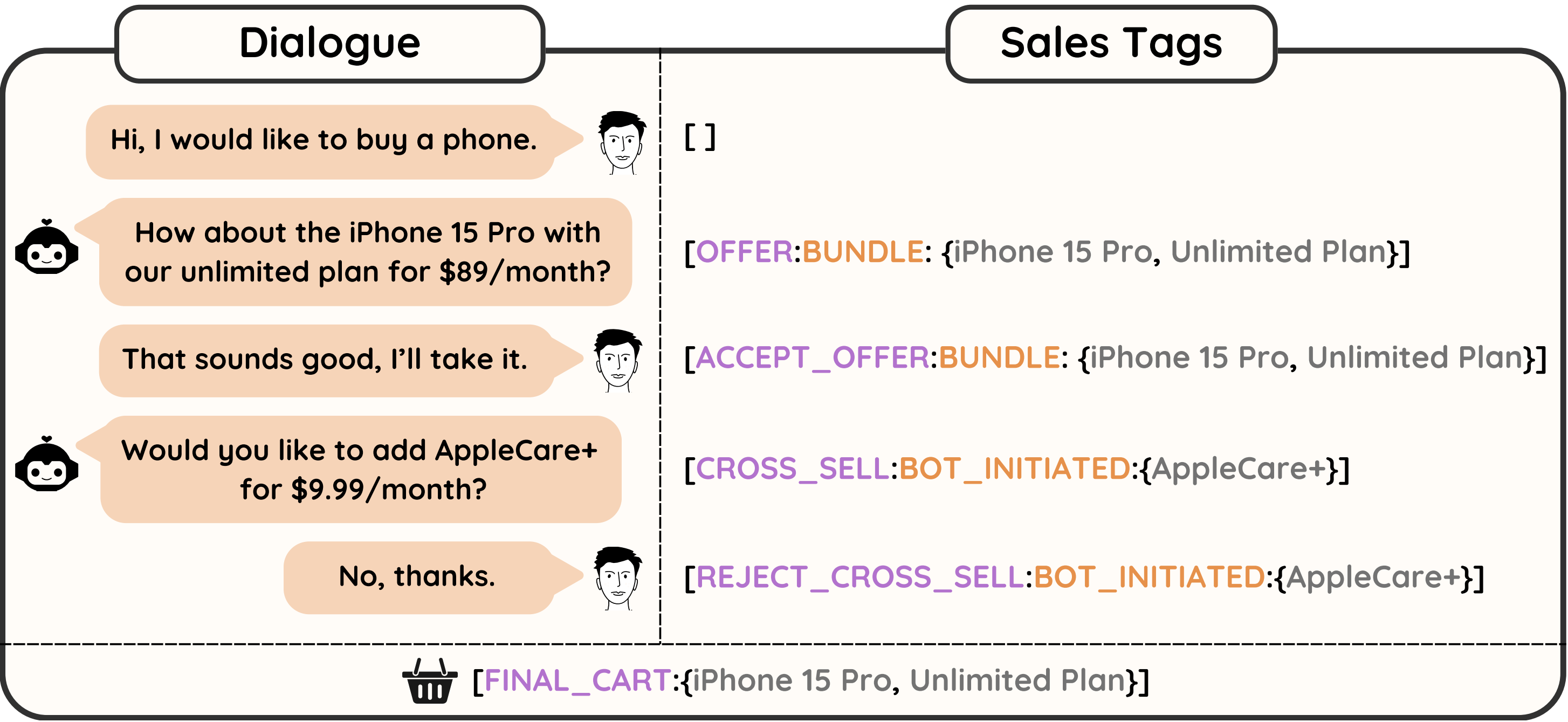
User satisfaction annotation

- 3 human annotators (104 dialogues)
- LLM-as-a-judge: Qwen-14B-1M (5200 dialogues)
- Alignment with the mean of human ratings: Spearman $\rho=0.78$



Sales tags annotation & Metrics

- Selected model: Claude 3.5 Sonnet v2

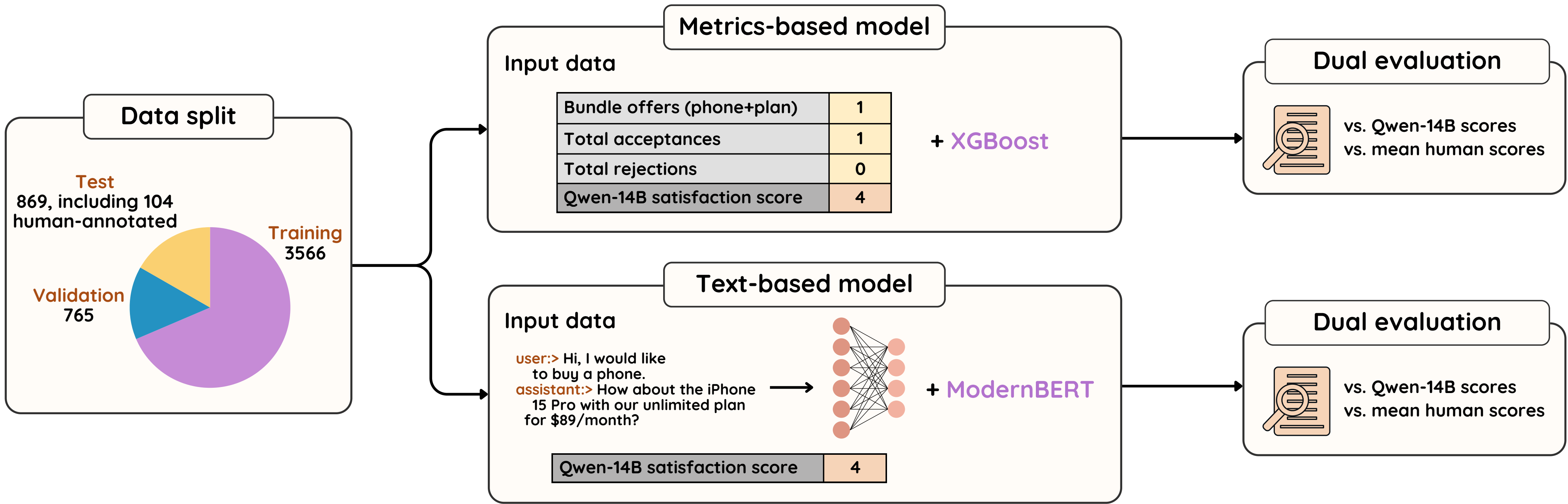


Metrics

Sales tags used for creating 120 interpretable metrics in 6 categories:

- **Basic**, e.g., counts of offers, acceptances, rejections.
- **Cart**, e.g., cart completion rate, final cart size, abandonment rate.
- **Conversion**, e.g., offer acceptance rates, cancellation patterns.
- **Flow**, e.g., negotiation patterns, sales density, customer journey progression.
- **Strategy**, e.g., product mix ratios, initiative rates, recovery success.
- **Timing**, e.g., turns to first offer, negotiation length, sales cycle duration.

Experiment: Metrics-based model vs. Text-based model



Results

Metric	Metrics-based model		Text-based model	
	Qwen-14B	Human	Qwen-14B	Human
MSE ↓	0.24	0.34	0.14	0.29
MAE ↓	0.40	0.42	0.28	0.41
R ²	0.87	0.78	0.92	0.81
ρ	0.82	0.75	0.89	0.75

Table 2. Regression results by model and test set (Qwen-14B: n=869, Human: n=104). Metrics-based: XGBoost (91 features), Text-based: ModernBERT.

Top ranked features from Metrics-based model

Rank	Feature	Definition	Category
1	Net conversion efficiency rate	$(A - C) / T$	Flow
2	Negotiation success rate	Proportion of offer-response cycles ending in acceptance	Flow
3	Num final cart items	Count of items in the final cart (F)	Cart-specific
4	Nbr turns	Total number of dialogue turns (T)	Dialogue structure
5	Customer decisiveness score	$A / (A + R + C)$	Flow
6	Post-rejection offer rate	Proportion of rejections (R) followed by a new offer	Strategy
7	Net overall conversion rate	Net acceptances (A - C) relative to all sales actions (O + CS)	Conversion
8	Adaptation score	Proportion of rejections or cancellations (R + C) followed by a new offer	Strategy
9	Offer rejection rate	Proportion of offers (O) explicitly rejected (R _o): R_o / O	Conversion
10	Phone-to-bundle ratio	P / B	Strategy
11	Mean user turn length	Average number of words per user turn	Dialogue structure
12	Stdev user turn length	Standard deviation of user turn lengths (words)	Dialogue structure
13	Rejection recovery success rate	Proportion of rejections (R) eventually followed by an acceptance	Strategy
14	Stdev turn length	Standard deviation of turn lengths across all speakers (words)	Dialogue structure
15	Has final purchase	$1[F > 0]$: 1 if the final cart is non-empty	Cart-specific

Table 3. Top 15 features ranked by mean absolute SHAP value on Qwen-14B test set. A = acceptances, C = cancellations, R = rejections, O = offers, CS = cross-sells, T = turns, F = final cart items, P = phone-only offers, B = bundle offers.

Insights for sales strategy

- **Agents should prioritize offer quality & variety over quantity.**
Successful conversion → higher satisfaction.
- **Persistence in negotiation matter.**
Continuing to make offers after rejection → higher negotiation success rates & higher satisfaction.
- **Engagement itself has value even without a purchase.**
Short failed conversations → lowest score, but longer conversations without a purchase can still maintain neutral to positive satisfaction.

Key findings

- Both models show comparable ranking ability vs. human judges.
- Satisfaction driven primarily by outcome-oriented features.
- Conversation process and sales strategy matter beyond outcomes:
 - Persistence after rejection associated with higher satisfaction.
 - Longer conversations without purchase can still yield neutral-to-positive satisfaction.
- Both models are highly reliable (<1% significant mispredictions); errors follow interpretable patterns:
 - **Overestimations:** agent ignored customer constraints despite prior acceptance.
 - **Underestimations:** helpful guidance without purchase still perceived positively.
- Metrics-based model provides full interpretability at marginal accuracy cost ($\Delta R^2 = 0.03-0.05$ vs. text-based), enabling actionable insights into sales strategy.