

Evaluation of Conversational Agents and Interpretable Satisfaction Modeling in Sales Dialogues

Justyna Gromada

Natalia Krawczyk



Orange Research
Shape the future and value innovation



Who we are?

Orange Research

led by Lyse Brillouet

**Orange
Research**

Shape the future
and value innovation

ACC Domain

Augmented Customers & Collaborators

led by Thierry Nagellen

CURIE Program

CUSomer-xR/aI Interplay for Individualized Experience

led by Michal Szczerbak



CURIE

#HumanCentric

#AgenticAI

#ResponsibleAI

#ConversationalAI

natural-in Project

Natural Interactions with Conversational Agents

led by Justyna Gromada

#ProactiveSalesBot

#Evaluation

#UserSimulator

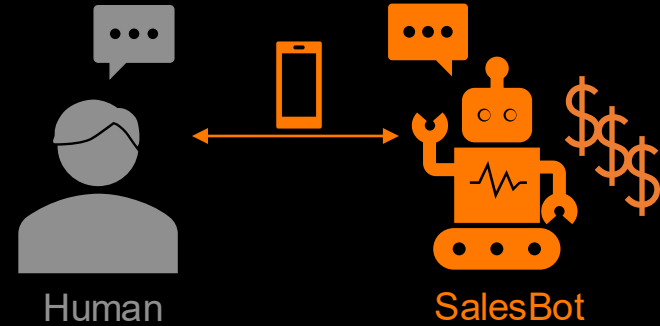
#Persona

#MultiParty

#Memory

Proactive SalesBot

- Chatbot helping customers find a smartphone
- R&D version
- Powered by LLM: GPT-4o-mini

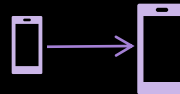


Proactive features:

- Cross-sell

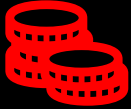


- Up-sell



The Challenge of Evaluating Conversational Agents

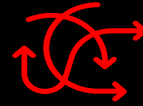
Human evaluation is the gold standard, but:



costly



time consuming



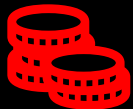
complicated



hard to scale

The Challenge of Evaluating Conversational Agents

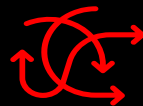
Human evaluation is the gold standard, but:



costly



time consuming



complicated



hard to scale

AI offers promising automated alternatives:



LLM-based user simulator

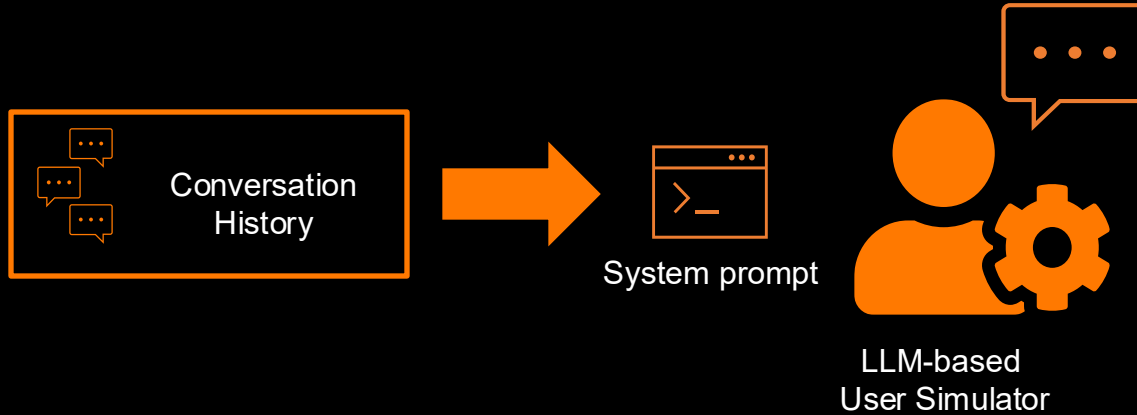
Pretends to be the customer



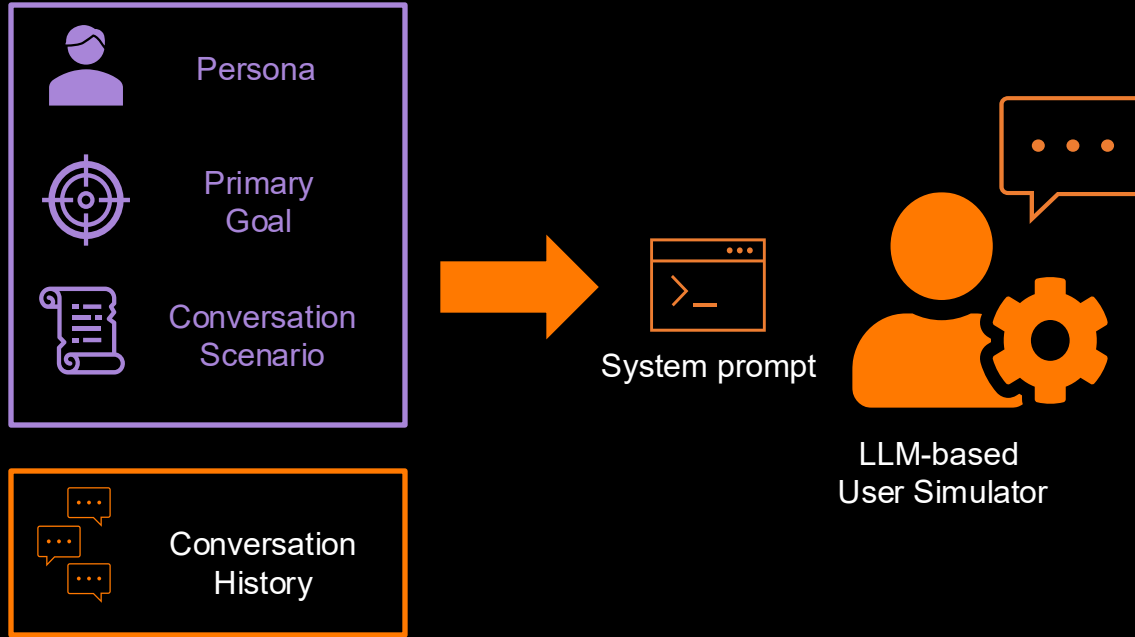
LLM-as-a-judge

Rates the conversation quality

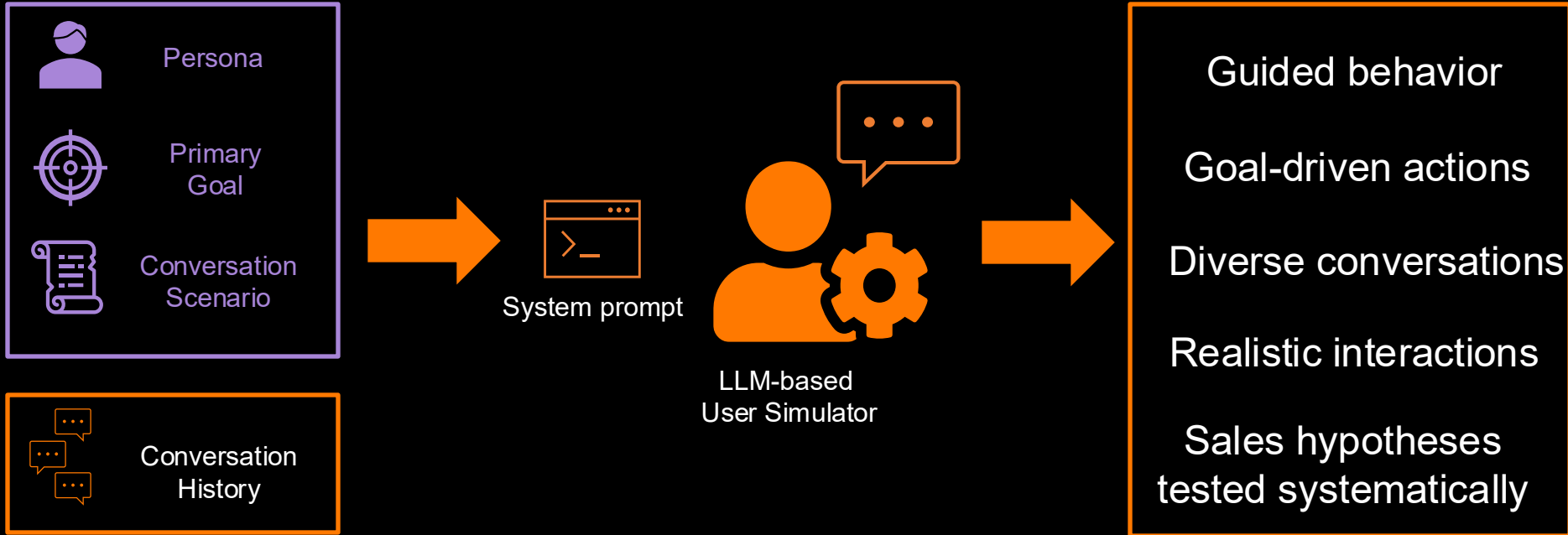
LLM-based User Simulator



Persona-driven LLM-based User Simulator



Persona-driven LLM-based User Simulator



Persona-driven LLM-based User Simulator

SalesBot Case



Persona

Orange Poland
customers

- Official market data
- Polish demographics
- Expert knowledge



Primary
Goal

Typical mobile
phone purchase



Conversation
Scenario

Based on:

- Assertiveness level
- Precision of needs

Persona-driven LLM-based User Simulator

SalesBot Case



Persona

Orange Poland customers

- Official market data
- Polish demographics
- Expert knowledge



Primary Goal

Typical mobile phone purchase



Conversation Scenario

Based on:

- Assertiveness level
- Precision of needs

Anthropic Claude 3.5 Sonnet v2:

- 100 personas representing Orange Poland customers
- 2 primary goals for each persona
- 4 conversational scenarios



PersonasForSalesbot



Exemplary dialogue



Name: **Jakub Nowicki**
Gender: **male**
Age: **23**
Occupation: **Software Developer**
Hobbies: **gaming, rock climbing, coding**
Income: **high**
Technology profile:
- Technology usage: **tech-savvy**
Telecom status:
- Current Orange client: **true**
- Customer_type: **business**



Goal ID: **PL_001_1**
Type: **phone upgrade**
Target user: **self**
Contract type: **postpaid**
Context: **As a software developer with high income, looking to upgrade from iPhone 14 Pro to latest model**



Assertiveness: **HIGH**
Precision of needs: **HIGH**



Hi

Hello! How can I assist you today? Are you looking for a new phone or a plan?



I'm interested in **upgrading my phone**. I'm currently on a business contract and I'd like to see what options are available for **the newest iPhone**.



Which subscription option would you like to choose? Here are the available options: 'choosing a new number', 'transitioning from nju to a subscription', 'transitioning from prepaid to a subscription', 'extending the contract', 'porting the number'.

I'm looking to **extend my current business contract**.



What is the maximum monthly installment you could afford?

...

No, I told you I want a phone with minimum 1TB storage



Our publications

Evaluating Conversational Agents with Persona-driven User Simulations based on Large Language Models: A Sales Bot Case Study

Justyna Gromada, Alicja Kasicka, Ewa Komkowska,
Łukasz Krajewski, Natalia Krawczyk, Bartosz Przybył,
Morgan Veyret, Lina Rojas-Barahona, Michał K. Szczerbak,
Orange Research



Evaluating Success in Proactive Sales Dialogues: Balancing Customer Experience and Business Outcomes

Justyna Gromada, Alicja Kasicka, Natalia Krawczyk, Michał Szczerbak
Orange Research



From Simulation to Reality: Validating Automated Evaluation of Conversational Agents

Justyna Gromada, Ekaterina Isakova, Joanna Brudnik, Alicja Kasicka,
Natalia Krawczyk, Bartosz Przybył, Łukasz Krajewski, Michał Szczerbak
Orange Research, Orange Poland



Interpretable Satisfaction Modeling in Sales Dialogues: A Metrics-based Approach Validated Against Text-based Model

Natalia Krawczyk, Alicja Kasicka, Bartosz Przybył, Justyna Gromada
Orange Research



Our publications

Evaluating Conversational Agents with Persona-driven User Simulations based on Large Language Models: A Sales Bot Case Study

Justyna Gromada, Alicja Kasicka, Ewa Komkowska,
Łukasz Krajewski, Natalia Krawczyk, Bartosz Przybył,
Morgan Veyret, Lina Rojas-Barahona, Michał K. Szczerbak,
Orange Research



Evaluating Success in Proactive Sales Dialogues: Balancing Customer Experience and Business Outcomes

Justyna Gromada, Alicja Kasicka, Natalia Krawczyk, Michał Szczerbak
Orange Research



From Simulation to Reality: Validating Automated Evaluation of Conversational Agents

Justyna Gromada, Ekaterina Isakova, Joanna Brudnik, Alicja Kasicka,
Natalia Krawczyk, Bartosz Przybył, Łukasz Krajewski, Michał Szczerbak
Orange Research, Orange Poland

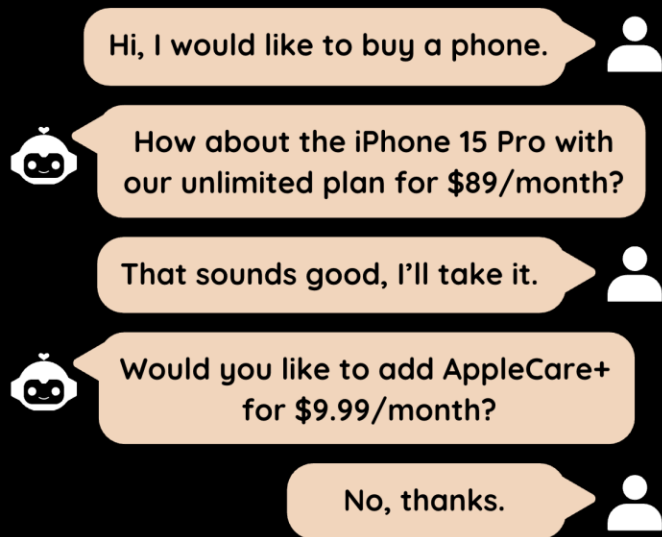


Interpretable Satisfaction Modeling in Sales Dialogues: A Metrics-based Approach Validated Against Text-based Model

Natalia Krawczyk, Alicja Kasicka, Bartosz Przybył, Justyna Gromada
Orange Research



Interpretable Satisfaction Modeling in Sales Dialogues: A Metrics-based Approach Validated Against Text-based Model



Why study satisfaction?

No purchase, but satisfied



Purchase, but dissatisfied



Task success $\neq$ customer satisfaction

Satisfaction reflects both what was achieved and how the interaction felt

Research questions



Can interpretable sales metrics predict customer satisfaction as well as a text-based transformer?



What influences customer satisfaction?

Simulated phone-sales dialogues

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 (100 P x 1 G x 4 S)
Total dialogues	5200

P: Personas, G: Goals, S: Scenarios

User satisfaction annotation



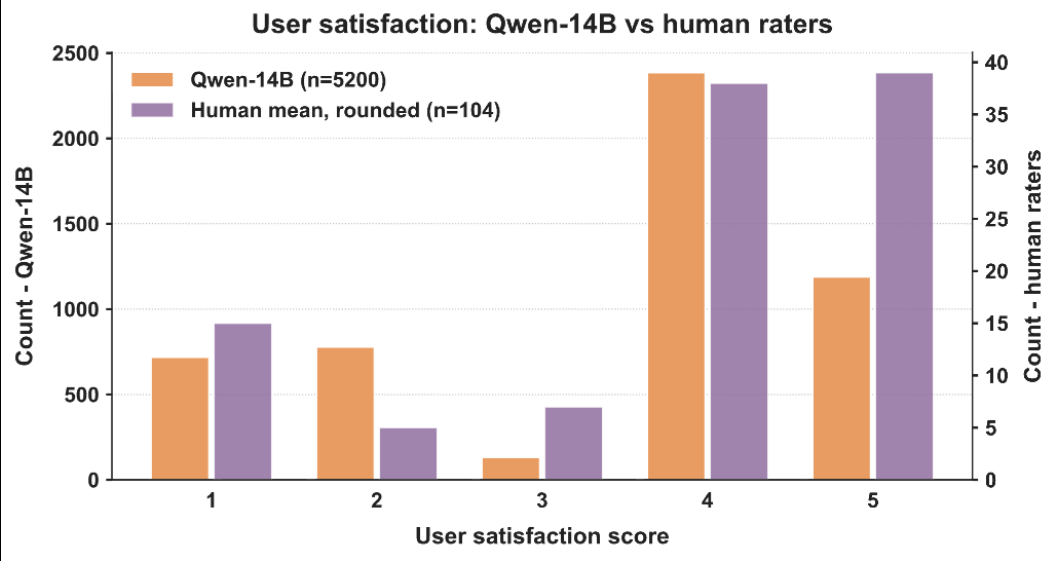
3 human annotators
104 dialogues



LLM-as-a-judge: Qwen-14B-1M
5200 dialogues

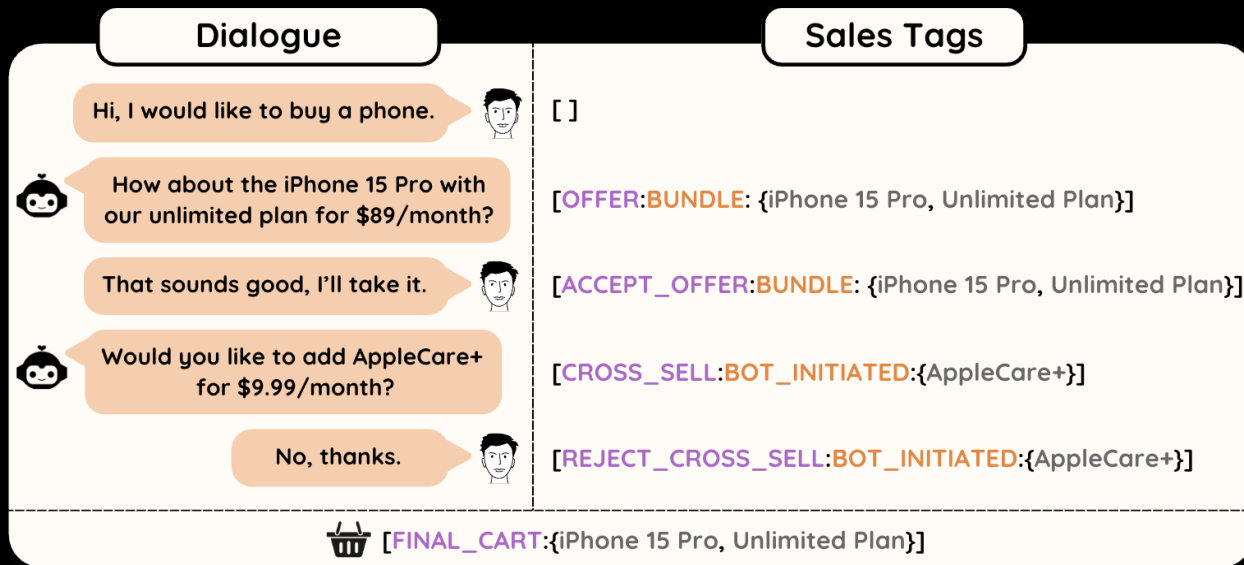
Alignment with the mean of human ratings: Spearman $\rho=0.78$

User satisfaction annotation



Sales tags annotation

Selected model: Claude 3.5 Sonnet v2

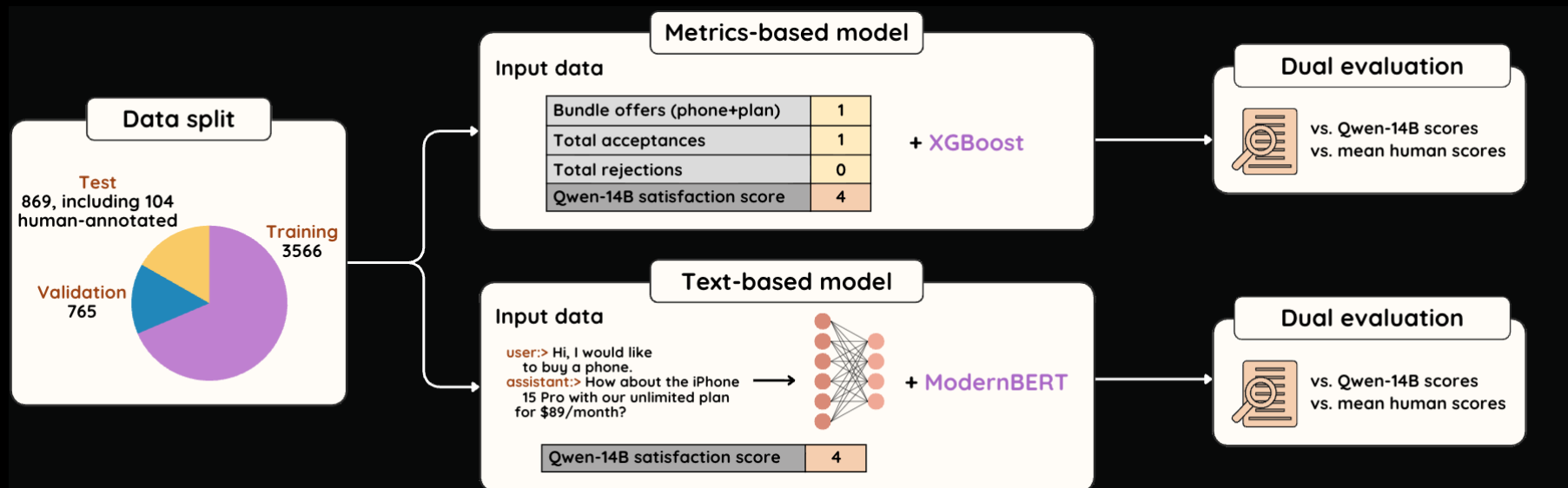


Metrics

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 and key findings

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

Qwen-14B: n = 869

Human: n = 104

Metrics-based: XGBoost (91 features)

Text-based: ModernBERT

Results and key findings

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

Qwen-14B: n = 869

Human: n = 104

Metrics-based: XGBoost (91 features)

Text-based: ModernBERT



Both models show comparable ranking ability vs. human judges

Results and key findings

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

Qwen-14B: n = 869

Human: n = 104

Metrics-based: XGBoost (91 features)

Text-based: ModernBERT



Both models show comparable ranking ability vs. human judges



Metrics-based model provides full interpretability at marginal accuracy cost ($\Delta R^2 = 0.03 - 0.05$), enabling actionable insights into sales strategy

Top ranked features from Metrics-based model

Rank	Feature
1	Net conversion efficiency rate
2	Negotiation success rate
3	Num final cart items
4	Nbr turns
5	Customer decisiveness score
6	Post-rejection offer rate
7	Net overall conversion rate
8	Adaptation score
9	Offer rejection rate
10	Phone-to-bundle ratio
11	Mean user turn length
12	Stdev user turn length
13	Rejection recovery success rate
14	Stdev turn length
15	Has final purchase

Top ranked features from Metrics-based model

Rank	Feature	Definition	Category
1	Net conversion efficiency rate	$(A - C)$ per dialogue turn: $(A - C) / T$	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

A = acceptances, C = cancellations, R = rejections, O = offers, CS = cross-sells, T = turns, F = final cart items, P = phone-only offers, B = bundle offers

Top ranked features from Metrics-based model

Rank	Feature
1	Net conversion efficiency rate
2	Negotiation success rate
3	Num final cart items
4	Nbr turns
5	Customer decisiveness score
6	Post-rejection offer rate
7	Net overall conversion rate
8	Adaptation score
9	Offer rejection rate
10	Phone-to-bundle ratio
11	Mean user turn length
12	Stdev user turn length
13	Rejection recovery success rate
14	Stdev turn length
15	Has final purchase



Satisfaction is primarily driven by outcome-oriented features

Insights for sales strategy

Agents should prioritize offer quality & variety over quantity



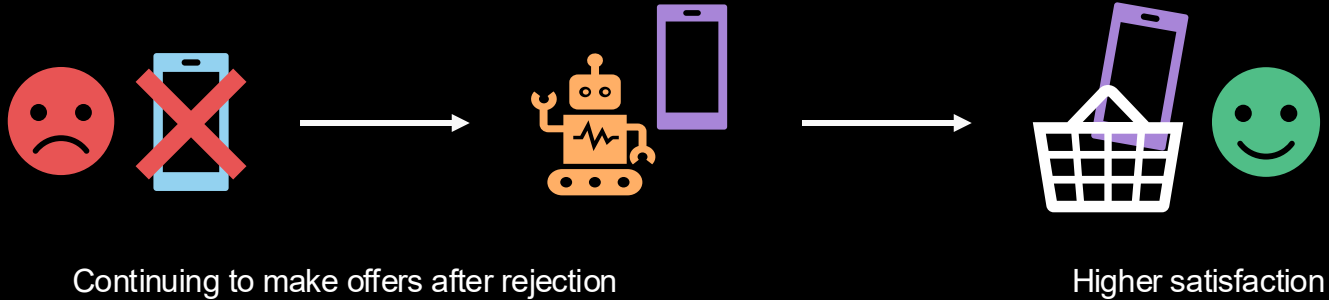
Successful conversation



Higher satisfaction

Insights for sales strategy

Persistent in negotiation matter



Insights for sales strategy

Engagement itself has value even without a purchase



Short failed conversations



Lowest satisfaction



Longer conversations



Higher satisfaction

Dziękujemy



justyna.gromada@orange.com
natalia1.krawczyk@orange.com



Orange Research
Shape the future and value innovation

