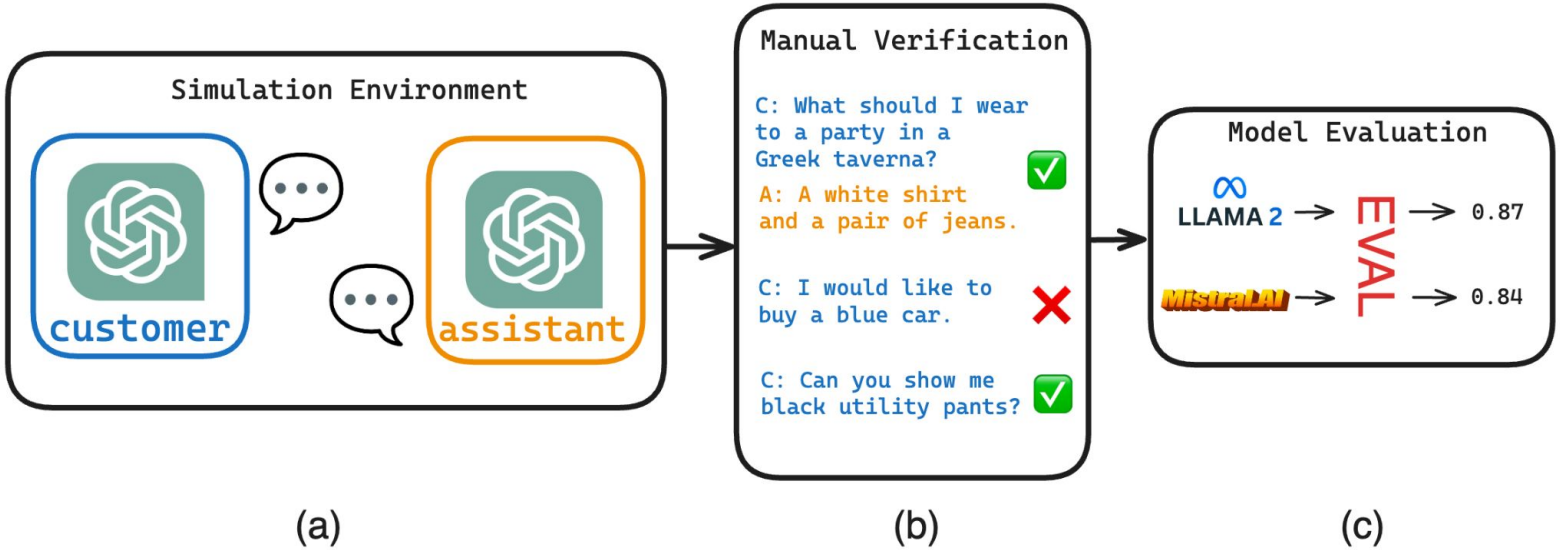
What should I wear to a party in a Greek taverna?

Evaluation for Conversational Agents in the Fashion Domain

Antonis Maronikolakis, Ana Peleteiro Ramallo, Weiwei Cheng, Thomas Kober
KDD workshop on Evaluation and Trustworthiness of Generative AI Models, 2024

zalando

Overview



A high-level overview of our methodology. We begin with (a) the generation of conversations through a simulation environment, which (b) we manually verify and (c) use to evaluate models.

Conversation Simulation

- Customer-Assistant agent interaction
- Customer is given instructions to shop for a particular *item* or *occasion*
- Conversations are recorded
- The customer agent is either LLM- or template-based

Customer

Assistant

Item Descriptions

- Fashion attributes are used for item descriptions:
  - Apparel (pants, trousers, shoes, jacket, coat, sweatshirt, hoodie and jeans)
  - Colour
  - Type (utility, athletic, formal, summer and urban)
  - Material
  - Fit
  - Brand (Converse, Nike, Carhartt, Pier One, Superdry, Levi's, Hugo Boss and PULL&BEAR)
  - Size
- Example: {*black, Converse, shoes, large}*

Occasions

- For the evaluation of ZA in open-ended scenarios
  - Hiking
  - Christmas Dinner
  - Football in winter
  - Rooftop summer party
  - Techno party in Berlin
  - Coffee house in Vienna
  - Party in Greek taverna
  - Sports in the summer
  - Eccentric timelord at book bazaar
- Note: Only used as input for LLM-based agents

Manual Verification

- Cleaning data for errors:
  - **Incorrect format** (customer agent not acting like a customer)
  - **Error messages**
  - **Incorrect customer behavior**
  - **Unfaithful** (incorrect details, accepting irrelevant item)
- For multilinguality, conversations are translated with GPT4 and verified

Manual Verification

- Cleaning data for errors:
  - **Incorrect format** (customer agent not acting like a customer)
  - **Error messages**
  - **Incorrect customer behavior**
  - **Unfaithful** (incorrect details, accepting irrelevant item)
- For multilinguality, conversations are translated with GPT4 and verified
- **3k messages, 1.3k conversations per language**

Example Conversation (party in greek taverna)

C: I am going to a party in a Greek taverna soon and I would like to buy some clothes for the occasion.

A: Are you looking for a specific type of outfit for the party in the Greek taverna? For example, a casual shirt and shorts, or something more formal like a linen suit?

C: I am looking for a more casual look

A: I found some casual outfit options for the party in the Greek taverna: *search(linen shirt, casual summer shorts, comfortable sandals)*

Evaluation Tasks

- **`AssistantEval`**: Evaluate capabilities of LLM to interface between customer and search engine
  - Customer is shopping for specific item
  - Customer is shopping for an occasion
  - *Conversations are replayed*
- **`QueryGenEval`**: Assuming the customer has given all the information needed, can our module produce relevant queries?
  - Full conversations are given to the model

Evaluation Details

- Customer agent is given description of an item:
  - BERTScore between output query and description
- Customer agent is given occasion
  - Semantic similarity between output query and occasion
  - ADA embeddings
  - Assignment with cosine similarity
  - Calculate percentage of correct assignments

Performance in `AssistantEval` (item descriptions)

Model	English	German	French	Greek
GPT-3.5 (I)	87.5	86.3	86.5	83.1
GPT-3.5 (II)	87.4	86.5	**86.8**	82.7
GPT-3.5 (III)	88.1	86.5	86.6	**83.5**
GPT-4	**88.5**	**86.9**	**86.8**	**83.5**

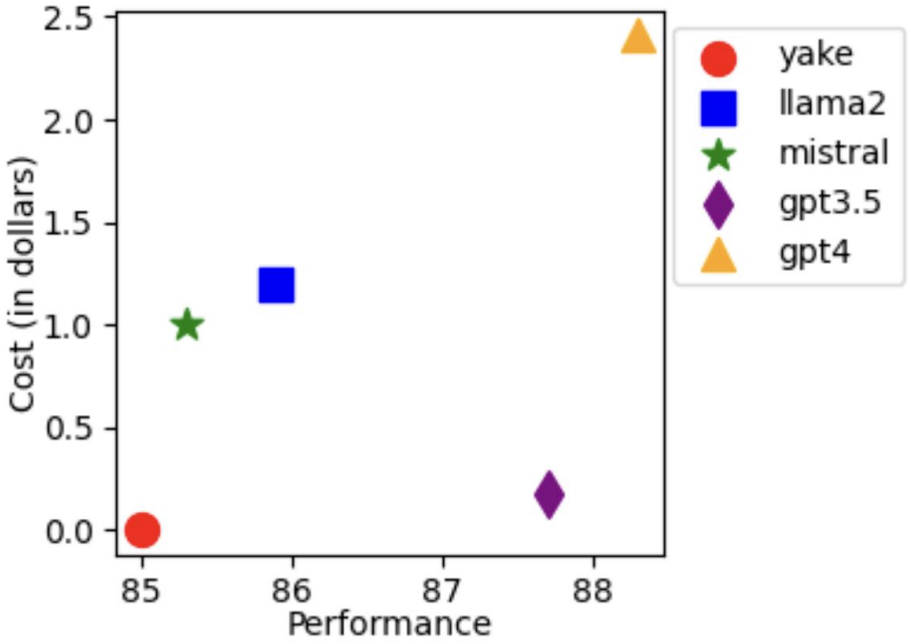
`AssistantEval` (item description) Qual. Analysis

Property	English	German	French
Colour	90.2	83.0	91.0
Type	35.3	21.2	34.9
Material	84.4	73.4	79.0
Fit	77.2	65.4	76.5
Brand	62.7	61.8	62.4
Apparel	72.2	67.1	72.3
Size	1.2	1.5	1.3

Performance in `QueryGenEval`

Model	English	German	French	Greek
Lorem	64.3	65.0	62.0	60.6
Popularity	72.5	74.1	75.2	73.1
Yake	85.0	84.0	84.2	82.4
Llama2	85.9	84.2	84.2	82.8
Mistral	85.3	84.0	84.1	82.3
GPT-3.5	87.7	86.2	**86.6**	**83.4**
GPT-4	**88.9**	**86.6**	86.5	83.3

QueryGenEval Cost Analysis



The answer to the question “what should I wear to a party in a Greek taverna” is, according to our best-performing agent, white shirts paired with loose jeans and comfortable shoes for dancing (no gloves for plate-breaking 😭)

Thank you