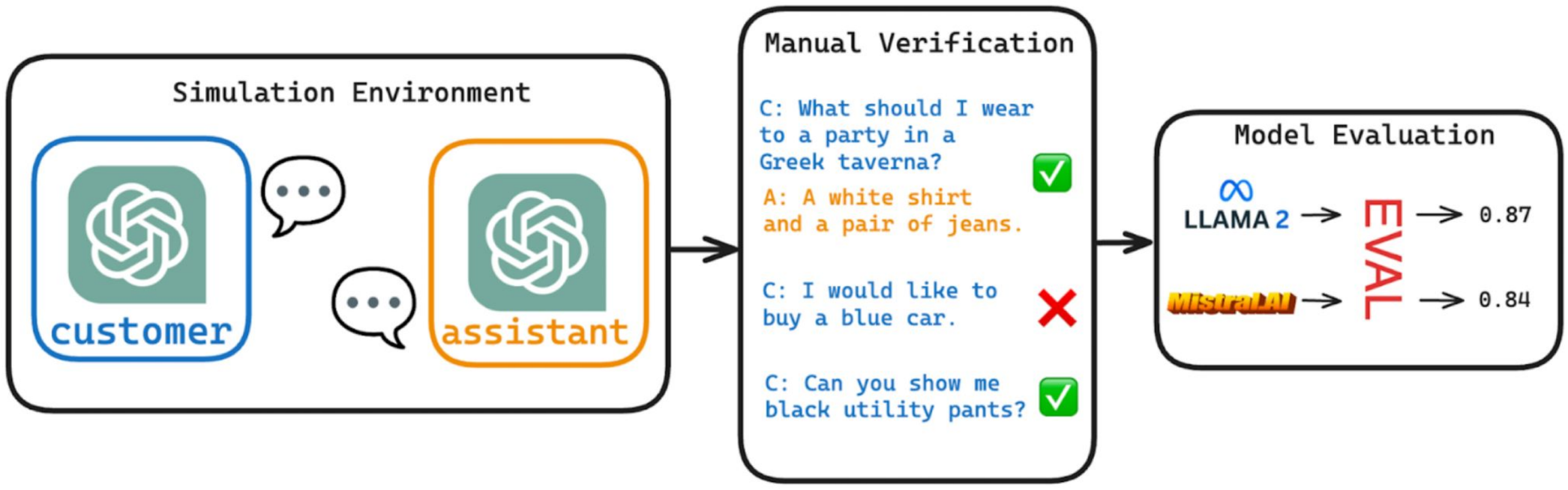
Evaluation of Conversational Agents in the Fashion Domain

Team Eliza

Antonis Maronikolakis, Ana Peleteiro Ramallo, Weiwei Cheng, Thomas Kober

We propose a methodology and dataset to evaluate the capabilities of conversational agents (such as the Zalando Assistant) to successfully interface between the customer and a search engine



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

GENERATED CONVERSATIONS

Description: black utility pants

C: Hi! I am looking for a pair of pants
A: I would love to help! Any particular style or colour?
C: Yes, I would like utility pants in black
A: Sure! Here are some items - [ITEMS]

Queries: black utility pants

EVALUATION TASKS

`AssistantEval`: Evaluation of agents to function as assistants between customer and search engine, both for open-ended scenarios and item-specific scenarios

`QueryGenEval`: Evaluation of agents to generate queries appropriate to the conversation and customer needs

RESULTS AND ANALYSIS

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

Figure: BERTScore F1 for **`QueryGenEval`**

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**

Figure: BERTScore F1 for **`AssistantEval`**

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

Figure: Qualitative analysis of GPT3.5 (III) performance in **`AssistantEval`** across fashion attributes