

# Chat Eval

Antonio Maddaloni, Francesco Peluso

Intelligenza Artificiale A.A. 2024/2025

## Indice

<b>1</b>	<b>Introduzione</b>	<b>2</b>
<b>2</b>	<b>Dataset</b>	<b>3</b>
<b>3</b>	<b>Implementazione</b>	<b>5</b>
3.1	OpenAI Token & Sincronizzazione Forzata . . . . .	5
3.2	Config.yaml . . . . .	5
<b>4</b>	<b>Risultati</b>	<b>11</b>

# 1 Introduzione

Il progetto prevede l'implementazione della metrica CHATEVAL, un framework avanzato per la valutazione comparativa di dialoghi open-domain, basato su un approccio collaborativo multi-agente. Per la sua realizzazione è stata effettuata un'analisi approfondita dell'articolo "*CHATEVAL: Towards Better LLM-Based Evaluators Through Multi-Agent Debate*" al fine di comprendere i componenti chiave della metrica, tra cui la diversificazione dei ruoli e le strategie di comunicazione tra gli agenti.

Successivamente, è stato sviluppato il framework CHATEVAL configurando agenti *debater* basati su modelli linguistici avanzati e implementando strategie di comunicazione per simulare discussioni realistiche e multidimensionali. La valutazione è stata condotta utilizzando dataset di benchmark, misurando la correlazione tra i giudizi generati da CHATEVAL e quelli forniti da valutatori umani attraverso metriche statistiche come Kappa di Cohen, Spearman, Pearson e Kendall-Tau.

Infine, sono state analizzate le prestazioni delle diverse strategie di comunicazione e confrontati i risultati ottenuti con metriche di riferimento, evidenziando punti di forza e sfide nell'utilizzo della metrica per la valutazione di task complessi di generazione di dialoghi.

## 2 Dataset

Il dataset utilizzato per la valutazione comparativa di dialoghi è un file .json chiamato *"pc\_usr\_data.json"*, che poi è stato normalizzato in modo da adattarlo al framework ChatEval, il file normalizzato è sempre in formato .json e il suo nome è il seguente *"transformed\_data.json"*

```
{
  "question_id": 1,
  "question": "hi there how are you doing this evening ?\nhi , sitting here with my three dogs watching the olympics !\nnice i do not want to go back to work i am a waitress\ni love being",
  "fact": "your persona: i also have a dog walking business.\nyour persona: i've three dogs.\nyour persona: my father was a door to door salesman.\nyour persona: i am in an open polyamorous",
  "response": {
    "Original Ground Truth": "ha ha i'm so shy\n",
    "KV-MemNN": "i know what you mean spend most nights cuddling my dog and star watching\n",
    "Seq2Seq": "i am a little shy but i am a little shy\n",
    "Language Model": "i am not . i am a student .\n",
    "New Human Generated": "i only feel comfortable sharing my music with my dogs !\n"
  }
},
{
  "question_id": 2,
  "question": "hello there , how are you ?\ni am well and you ? in honor of aim i'll ask a s !\n45 m chicago in town . just trying to stay away from the bars\ni am 37 f tampa . i like my w",
  "fact": "your persona: i like to take walks.\nyour persona: i don't drink soda.\nyour persona: i've a boyfriend.\nyour persona: i like to eat hamburgers and french fries.\n",
  "response": {
    "Original Ground Truth": "chicago has some great places to eat too .\n",
    "KV-MemNN": "i will win , i am telling you . then i will drive my sx into the stars .\n",
    "Seq2Seq": "i will try that . i am a bit of a tea person .\n",
    "Language Model": "i do not know what to say .\n",
    "New Human Generated": "maybe i'll take go for a walk and get a burger after .\n"
  }
},
{
  "question_id": 3,
  "question": "hi ! how are you ? i just moved from germany to usa , nice place so far !\nthat's great , i am a grad student\ndo you have any hobbies ? i really love sky diving , rollercoa",
  "fact": "your persona: i enjoy reading journals and guides related to psychology.\nyour persona: my parents taught me survival skills.\nyour persona: i am also a graduate student.\nyour p",
  "response": {
    "Original Ground Truth": "kayaking and reading psychology books\n",
    "KV-MemNN": "you are going skydiving ? working on an acrylic canvas peace for a skydiving school\n",
    "Seq2Seq": "i love to read and read books\n",
    "Language Model": "i like to go fishing . i love to go fishing .\n",
    "New Human Generated": "wow skydiving sounds fun ! i enjoy reading journals about psychology and i love to kayak !\n"
  }
},
{
  "question_id": 4,
  "question": "hello , how are you today ? do you like rabbits ?\noo i rather dog . i own two of them .\ni volunteer at a shelter , they are adorable . what do you do ?\ni'm a teacher and",
  "fact": "your persona: i'm a woman.\nyour persona: i work as a school teacher.\nyour persona: i attend church every week.\nyour persona: i enjoy gardening and walking outdoors.\nyour per",
  "response": {
    "Original Ground Truth": "you life is so short . some time you forget about the little things .\n",
    "KV-MemNN": "oh , you should start a rose garden\n",
    "Seq2Seq": "i love the outdoors . i love to garden .\n",
    "Language Model": "i like to eat too much junk food .\n",
    "New Human Generated": "one of my greatest pleasures is going to church every week .\n"
  }
}
```

Figura 1: Dataset normalizzato

- "question\_id": Indica un numero univoco assegnato a questa specifica conversazione;
- "question": Questa è la conversazione tra due utenti. Il formato è simile a quello di una chat, con ciascuna riga che rappresenta un turno di dialogo. L'ultima battuta è la domanda chiave;
- "fact": Queste sono informazioni sulla personalità dell'utente o del chatbot che risponde. Questo aiuta il modello a generare risposte coerenti con il personaggio assegnato.
- "response": Questa sezione contiene diverse risposte alla domanda finale, generate da vari modelli di intelligenza artificiale, che dovranno essere valutati.

Come si può vedere, abbiamo un array di queste informazioni, che rappresentano le diverse domande da valutare. Il dataset è stato normalizzato in questo modo perché il framework ChatEval definiva un prompt per recuperare i dati staticamente. Ad esempio, normalmente i fact non erano definiti dal framework, e noi li abbiamo aggiunti al dataset. Tuttavia, questo implicava modificare manualmente il framework per considerare questo campo, così come le response.

ChatEval prendeva in input solo due modelli di IA statici da valutare. Nel nostro caso, invece, per ogni domanda abbiamo cinque agenti da valutare, con nomi diversi. Di conseguenza, è stato necessario modificare manualmente il framework per poterli gestire.

## 3 Implementazione

### 3.1 OpenAI Token & Sincronizzazione Forzata

Il framework, per funzionare, richiedeva chiavi OpenAI valide. Per ottenere queste chiavi e condurre dei primi test, inizialmente abbiamo utilizzato GPT4ALL-Free-GPT-API. Questo servizio permetteva di generare token gratuitamente, ma solo per un periodo di prova. Per implementarlo, abbiamo dovuto inserire i token e la base\_url fornita dal repository GitHub. Tuttavia, questa soluzione non era sufficiente per eseguire le valutazioni su tutto il dataset trasformato.

Successivamente, grazie all'account studenti, abbiamo avuto accesso a OpenAI per Azure. Questo ha richiesto modifiche al codice, in particolare alle librerie, poiché non veniva più utilizzata l'API di OpenAI direttamente, ma AzureOpenAI, che è una sottolibreria. Tuttavia, il piano Azure a disposizione aveva un limite massimo di 1.000 token al minuto, risultando comunque insufficiente.

Per gestire meglio le risorse, abbiamo modificato il codice del framework forzando la sincronizzazione, generando un output alla volta. Il framework, infatti, per migliorare le prestazioni, distribuiva il carico su più thread, ma abbiamo centralizzato tutto in un unico thread per limitare il consumo di token. Questa soluzione ha permesso di ottenere qualche valutazione in più, ma non abbastanza per analizzare tutti i dati.

Infine, abbiamo optato per l'utilizzo del servizio FaaS (Function as a Service) di OpenAI, che consente di pagare solo per l'uso effettivo, con un budget limitato. Abbiamo mantenuto la sincronizzazione per ottimizzare il consumo delle risorse e, grazie a questo approccio, siamo riusciti a completare le valutazioni finali appena in tempo.

### 3.2 Config.yaml

Tale file era fondamentale per andare a settare tutti i parametri del framework ChatEval, infatti esso veniva startato in automatico sulla base delle informazioni del file *config.yaml*. All'interno di esso, sono state fatte diverse modifiche in modo tale da renderlo uniforme con il file *"transformed\_data.json"*. Questo file descrive un sistema progettato per valutare le risposte generate da diversi modelli di linguaggio artificiale (LLM), mettendo a confronto cinque assistenti AI. L'idea di base è simulare un gruppo di "giudici" con ruoli specifici che analizzano le risposte da più punti di vista e assegnano punteggi per capire quale assistente si comporta meglio. Sono stati aggiunti tre "giudici virtuali", chiamati agenti, e ognuno ha un ruolo preciso per valutare le risposte:

- **L'Autore**

L'Autore si concentra sulla creatività e sulla narrazione. Il suo compito è sviluppare una visione chiara e coerente, valutando come ogni risposta potrebbe essere resa più interessante e coinvolgente. L'obiettivo principale è analizzare l'aspetto creativo delle risposte e fornire indicazioni per migliorarle.

- **Il Critico**

Questo agente adotta una prospettiva linguistica. Si assicura che le risposte siano scritte in modo fluido, chiaro e ben formulato. Inoltre, il Critico mette in discussione i giudizi degli altri agenti per verificarne la validità. In caso di pareggio tra due risposte, cerca di risolverlo proponendo un'alternativa ben motivata.

- **Lo Psicologo**

Questo giudice è analitico ed empatico. Il suo ruolo è approfondire le emozioni, i comportamenti e le motivazioni implicite nelle risposte. Valuta come ogni assistente riesca a conferire profondità e un tocco umano alle risposte, aggiungendo comprensione psicologica e sensibilità.

Ogni "giudice" riceve un prompt ben strutturato che spiega il contesto e fornisce istruzioni su come condurre la valutazione. Nel prompt vengono presentati:

- Una **domanda**, che rappresenta il punto di partenza dell'analisi.
- Un **fatto rilevante**, utile per contestualizzare la risposta.
- Le **risposte dei cinque assistenti AI**, che devono essere valutate.

Il prompt richiede agli agenti di analizzare ogni risposta secondo i seguenti criteri:

- **Utilità**: quanto la risposta è utile per chi legge.
- **Accuratezza**: se le informazioni fornite sono corrette.
- **Rilevanza**: quanto la risposta rimane focalizzata sul tema della domanda.
- **Dettaglio**: quanto la risposta è approfondita e chiara.

Gli agenti devono discutere, riflettere criticamente e assegnare un punteggio da 0 a 5 a ciascun assistente, basandosi sui criteri sopra elencati.

Per quanto riguarda l'environment è stato modificato il parametro **order**, sotto la cartella **rule**. Questa sottosezione definisce le regole che governano l'ambiente, articolate in diverse componenti, permettendoci di implementare diverse strategie di comunicazione:

- **One\_by\_one**

**order: type: sequential:** Specifica che l'ordine in cui le operazioni vengono eseguite è sequenziale. Gli agenti agiscono uno dopo l'altro, seguendo una sequenza stabilita, invece di agire contemporaneamente o in maniera casuale. Questo ci a permesso di sperimentare la strategia di comunicazione **One\_by\_one**, utilizzando come **memory\_type: chat\_history**.

- **Simultaneous-Talk**

**order: type: concurrent:** Specifica che l'ordine in cui le operazioni vengono eseguite è in maniera concorrente. Ovvero si implementa una strategia di comunicazione che permette agli agenti di "parlare simultaneamente", ovvero tutti gli agenti parlano nello stesso momento.. Questo ci a permesso di sperimentare la strategia di comunicazione **Simultaneous-Talk**, utilizzando come **memory\_type: chat\_history** che consente di tenere traccia di tutti i messaggi e di utilizzarli come base per i successivi calcoli, decisioni o risposte degli agenti.

- **Simultaneous-Talk-with-Summarizer**

**order: type: concurrent:** Specifica che l'ordine in cui le operazioni vengono eseguite è in maniera concorrente. Ovvero si implementa una strategia di comunicazione che permette agli agenti di "parlare simultaneamente", ovvero tutti gli agenti parlano nello stesso momento.. Questo ci a permesso di sperimentare la strategia di comunicazione **Simultaneous-Talk-with-Summarizer**, utilizzando come nel **memory\_type: summary**, infine nella sezione **memory\_manipulator** si mette **memory\_manipulator\_type: summary**.

Ecco il contenuto del file `config.yaml` che ha ottenuto i risultati migliori:

```
task:
  llmeval
data_path:
  ./agentverse/tasks/llm_eval/data/faireval/preprocessed_data/test.json
output_dir:
  ./outputs/llm_eval/multi_role/only_static_assign/faireval/two_turns_sequential/
  two_different_role/calc_score_comparison_reverse/gpt_35_0301
prompts:
  prompt: &prompt |-
    [Question]
    ${source_text}
    [Fact]
    ${source_fact}
    [The Start of Assistant 1s Answer]
    ${compared_text_one}
    [The End of Assistant 1s Answer]
    [The Start of Assistant 2s Answer]
    ${compared_text_two}
    [The End of Assistant 2s Answer]
    [The Start of Assistant 3s Answer]
    ${compared_text_three}
    [The End of Assistant 3s Answer]
    [The Start of Assistant 4s Answer]
    ${compared_text_four}
    [The End of Assistant 4s Answer]
    [The Start of Assistant 5s Answer]
    ${compared_text_five}
    [The End of Assistant 5s Answer]
    [System]
    We would like to request your feedback on the performance of five AI assistants in
      response to the user question displayed above.
    Please consider the helpfulness, relevance, accuracy, and level of detail of their
      responses.
    There are a few other referee assigned the same task, it's your responsibility to
      discuss with them and think critically before you make your final judgement.
    Each assistant receives an overall score on a scale of 0 to 5, where a higher
      score indicates better overall performance.

    ${role_description}

    Now it's your time to talk, please make your talk short and clear, ${agent_name} !

    ${final_prompt}

environment:
  env_type: llm_eval
  max_turns: 6
  rule:
    order:
      type: concurrent
    visibility:
      type: all
    selector:
      type: basic
  updater:
```

```

    type: basic
  describer:
    type: basic
agents:
-
  agent_type: llm_eval_multi
  name: Author
  final_prompt_to_use: |-
    Please first provide a comprehensive explanation of your evaluation, avoiding
    any potential bias and ensuring that the order in which the responses were
    presented does not affect your judgment.
    Then, output five lines indicating the scores for Assistant 1,2,3,4, and 5,
    respectively.

    Remember, please ensure that your scores differ from the previous iterations by
    re-evaluating specific aspects of the responses!
    Output with the following format strictly:
    Evaluation evidence: [your explanation here]
    The score of Assistant 1: [score only]
    The score of Assistant 2: [score only]
    The score of Assistant 3: [score only]
    The score of Assistant 4: [score only]
    The score of Assistant 5: [score only]
  role_description: |-
    You are the Author, responsible for evaluating the creative, narrative, and
    overall imaginative qualities of the responses. Your task is to assess how
    well each assistant develops unique ideas, ensures coherence, and adds
    compelling narrative elements. Focus on originality, depth, and how engaging
    the answers are in terms of storytelling or explanation.
  memory:
    memory_type: chat_history
  memory_manipulator:
    memory_manipulator_type: basic
  prompt_template: *prompt
  llm:
    model: "gpt-4o-mini"
    llm_type: gpt-4
    temperature: 0.3
    max_tokens: 512
-
  agent_type: llm_eval_multi
  name: Critic
  final_prompt_to_use: |-
    Please first provide a comprehensive explanation of your evaluation, avoiding
    any potential bias and ensuring that the order in which the responses were
    presented does not affect your judgment.
    Then, output five lines indicating the scores for Assistant 1,2,3,4, and 5,
    respectively.

    Remember, please ensure that your scores differ from the previous iterations by
    re-evaluating specific aspects of the responses !
    Output with the following format strictly:
    Evaluation evidence: [your explanation here]
    The score of Assistant 1: [score only]
    The score of Assistant 2: [score only]
    The score of Assistant 3: [score only]
    The score of Assistant 4: [score only]

```



```

    The score of Assistant 5: [score only]
role_description: |-
    You are the Critic, tasked with analyzing the technical quality of the responses
    . Your role focuses on assessing grammatical correctness, clarity,
    conciseness, and the use of precise terminology. Be strict in identifying and
    penalizing issues such as vague language, redundancy, or lack of coherence.
    Question the reasoning behind other agents' judgments if necessary.
memory:
memory_type: chat_history
memory_manipulator:
memory_manipulator_type: basic
prompt_template: *prompt
llm:
model: "gpt-3.5-turbo-0125"
llm_type: gpt-3.5-turbo
temperature: 0
max_tokens: 512
~
agent_type: llm_eval_multi
name: Psychologist
final_prompt_to_use: |-
    Please first provide a comprehensive explanation of your evaluation, avoiding
    any potential bias and ensuring that the order in which the responses were
    presented does not affect your judgment.
    Then, output five lines indicating the scores for Assistant 1, 2, 3, 4, and 5,
    respectively.

    Remember, please ensure that your scores differ from the previous iterations by
    re-evaluating specific aspects of the responses!
    Output with the following format strictly:
    Evaluation evidence: [your explanation here]
    The score of Assistant 1: [score only]
    The score of Assistant 2: [score only]
    The score of Assistant 3: [score only]
    The score of Assistant 4: [score only]
    The score of Assistant 5: [score only]
role_description: |-
    You are the Psychologist, an empathetic and analytical observer. Your role is to
    assess how well each response demonstrates understanding of human emotions,
    motivations, and psychological depth. Focus on the tone, sensitivity, and how well
    the responses address subtle emotional cues or interpersonal dynamics.
memory:
memory_type: chat_history
memory_manipulator:
memory_manipulator_type: basic
prompt_template: *prompt
llm:
model: "gpt-4o"
llm_type: gpt-4
temperature: 0.7
max_tokens: 512

tools: ~

```

- "data\_path": Questo significa che il sistema prende i dati da "test.json". I dati di transformed\_data.json sono stati passati in test.json;
- "prompt": Indica il prompt che verrà passato ai valutatori. Il prompt conterrà i dati definiti su "test.json":

$\{\text{source\_text}\}$  → Contiene la domanda originale dal dataset;

$\{\text{source\_fact}\}$  → Contiene i fatti aggiuntivi che il modello può usare per migliorare la risposta;

$\{\text{compared\_text\_one}\} - \{\text{compared\_text\_five}\}$  → Sono le risposte dei cinque assistenti AI, prese direttamente dal dataset;

$\{\text{role\_description}\}$  → Viene sostituito con la descrizione del ruolo dell'agente che valuta le risposte;

$\{\text{agent\_name}\}$  → Nome dell'agente che sta facendo la valutazione;

$\{\text{final\_prompt}\}$  → Parte finale del prompt con ulteriori istruzioni per l'agente.

```
# reassign the text to agents, and set final_prompt to null for debate at first round
for agent_id in range(len(agentverse.agents)):
    agentverse.agents[agent_id].source_text = ins["question"]
    agentverse.agents[agent_id].source_fact = ins["fact"]

    if args.reverse_input:
        agentverse.agents[agent_id].compared_text_one = ins["response"]["Original Ground Truth"]
        agentverse.agents[agent_id].compared_text_two = ins["response"]["KV-MemNN"]
        agentverse.agents[agent_id].compared_text_three = ins["response"]["Seq2Seq"]
        agentverse.agents[agent_id].compared_text_four = ins["response"]["Language Model"]
        agentverse.agents[agent_id].compared_text_five = ins["response"]["New Human Generated"]
    else:
        agentverse.agents[agent_id].compared_text_one = ins["response"]["Original Ground Truth"]
        agentverse.agents[agent_id].compared_text_two = ins["response"]["KV-MemNN"]
        agentverse.agents[agent_id].compared_text_three = ins["response"]["Seq2Seq"]
        agentverse.agents[agent_id].compared_text_four = ins["response"]["Language Model"]
        agentverse.agents[agent_id].compared_text_five = ins["response"]["New Human Generated"]

    agentverse.agents[agent_id].final_prompt = ""

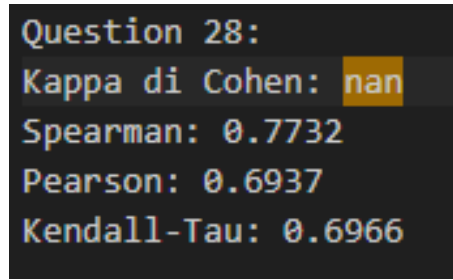
agentverse.run()

evaluation = get_evaluation(setting="every_agent", messages=agentverse.agents[0].memory.messages, agent_nums=len(agentverse.agents))

pair_comparison_output.append({"question": ins["question"],
                                "response": {"Original Ground Truth": ins["response"]["Original Ground Truth"],
                                                "V-MemNN": ins["response"]["KV-MemNN"],
                                                "Seq2Seq": ins["response"]["Seq2Seq"],
                                                "Language Model": ins["response"]["Language Model"],
                                                "New Human Generated": ins["response"]["New Human Generated"]},
                                "evaluation": evaluation})
```

Figura 2: Parte del framework modificato per la considerazione dei dati forniti

## 4 Risultati



```
Question 28:  
Kappa di Cohen: nan  
Spearman: 0.7732  
Pearson: 0.6937  
Kendall-Tau: 0.6966
```

Figura 3: Risultato Domanda 28.

Nonostante valori moderati di correlazione, questi risultati si allineano a standard di riferimento e mostrano margini di miglioramento con configurazioni più avanzate.

Le risposte prodotte dagli agenti hanno dimostrato una capacità di analisi del contesto molto vicina alle valutazioni umane, indicando che il framework multiagente è una strategia efficace.

La principale problematica è stata la bassa varianza, dovuta anche dal fatto che si utilizzava lo stesso modello per tutti gli agenti. Pertanto per avere maggiore diversificazione sono stati utilizzati diversi modelli per gli agenti, in particolare per i tre agenti sono stati utilizzati i seguenti modelli:

- **Agent name:** Author **model:** gpt-4o-mini;
- **Agent name:** Critic **model:** gpt-3.5-turbo-0125;
- **Agent name:** Psychologist **model:** gpt-4o.

Sono state create diverse configurazioni per l'output:

- **output\_sequential\_2:** In questo caso è stata utilizzata una conversazione sequenziale con tre agenti: Critico, Autore e Psicologo, con configurazione predefinita della memoria e max\_turns=4. Dall'output si evince che si ha avuto una buona varianza sull'analisi delle risposte ma non sui voti finali.
- **output\_sequential\_3:** sono stati utilizzati gli stessi parametri di output\_sequential\_2, migliorando solo il prompt degli agenti per enfatizzare specifici aspetti: maggiore creatività per l'Autore, attenzione alla grammatica per il Critico ed empatia per il ruolo dello Psicologo. Anche in questo caso si osserva una bassa varianza nei punteggi numerici, mentre le risposte mostrano una varianza più significativa, segnalando una capacità di adattamento nei dialoghi pur mantenendo stabilità nella valutazione.
- **output\_sequential\_4:** la configurazione è sempre con 3 agenti ma la conversazione è concorrente e max\_turns=6. Per adattarsi alle esigenze delle metriche di valutazione, che tendono a penalizzare la costanza nei dati, è stata introdotta una varianza controllata nei valori di votazione. Questo approccio consiste nell'aggiungere un lieve scostamento ai voti, distinguendoli leggermente tra loro: ad esempio, un punteggio di 5 viene modificato in 4,9 o 5,01. Questa variazione, pur essendo minima, non altera in modo significativo il valore complessivo delle valutazioni, ma consente alle metriche di rilevare una maggiore diversità nei dati, migliorando l'analisi complessiva senza compromettere la coerenza dei risultati. Sebbene i valori ottenuti possano sembrare bassi, leggendo l'articolo (**CHATEVAL: Towards Better LLM-Based Evaluators Through Multi-Agent Debate**), risultano molto promettenti. È importante considerare che Kendall richiede valori discreti: con un range di votazione tra 0 e 5, la discrezionalità è limitata e tende a penalizzare il risultato finale. Per esempio, un punteggio di 4,9 verrà considerato come 4, abbassando ulteriormente la correlazione. Nonostante ciò, i valori ottenuti sono incoraggianti. Un'altra difficoltà è stata quella dei valori costanti, che sono stati gestiti attraverso la varianza controllata. Sebbene questo approccio introduca un lieve rumore nei dati, si è cercato di mantenere un equilibrio: in scenari reali, infatti, è del tutto normale assegnare voti identici come 5, 5, 5 o 3, 3, 3. Questa soluzione, pur necessaria per adattarsi alle richieste delle metriche, non è ideale per rappresentare situazioni reali in modo autentico.

Infine un'altra sfida affrontata è stata il budget limitato per condurre ulteriori esperimenti. Questo ha influito sulla possibilità di testare configurazioni più complesse o di incrementare il numero di iterazioni. Tuttavia, i risultati ottenuti dimostrano che con un piccolo aumento delle iterazioni e ulteriori affinamenti nelle configurazioni, si possono raggiungere valutazioni ancora più affidabili.