Simulating Human Conversations: An LLM-based Approach to Synthetic Dialogue Generation and Evaluation

Gianmarco Cornacchia (2575307c)

COMPSCI5082P (40 Credits) - March 28, 2025

ABSTRACT

The development of high-quality conversational datasets is essential to the advancement of AI systems that rely on natural language. Collecting such data through human conversations is expensive, high-effort and hard to scale. Moreover, there is a lack of accessible tools to synthetically generate conversational data in a customisable and scalable way. This paper introduces a configurable, open-source conversation simulation platform that supports the use of small, freely available Large Language Models (LLMs) to generate customised, high-quality synthetic conversational datasets without incurring generation costs. An LLM-as-a-judge evaluation pipeline is proposed to assess the quality of the generated conversations and a human evaluation study is conducted to validate the reliability of such an evaluation pipeline. Results show that while LLM judges backed by small general-purpose open-source models tend to poorly correlate with human judgement, a fine-tuned evaluation model (Atla Selene Mini) demonstrates statistically significant positive alignment with human ratings for consistency and relevance. Interpreting the statistically significant LLM-as-a-judge evaluations and human evaluations it was found that conversations generated by small open-source LLMs can be of high quality, improving as the size of the backbone model behind the conversation simulator increases.

1. INTRODUCTION

Large-scale high-quality conversational datasets are crucial to the development of AI systems that involve natural language [12], however most existing conversational datasets lack the scale and specificity to be suitable for training dialogue systems [31]. Crafting new conversational datasets from human conversations is a complicated and expensive process [29], paving the way for synthetic data generation.

With the recent advancements in the field of generative AI brought by Large Language Models (LLMs), the use of LLMs for synthetic dialogue data generation is becoming increasingly common, however there is a lack of accessible specific tools to create conversational datasets in a customisable and scalable way.

Simulating patient-doctor dialogues for the training of medical dialogue systems [6], generating music recommendation dialogue for a music recommender system [15], and studying conversations between LLM agents for social interaction research [42] are only some examples of the wide range of specific domains who would benefit from an accessible general-purpose tool that allows the generation of synthetic conversational data in a tailored and scalable way when real-world data is scarce or difficult to obtain [29, 31].

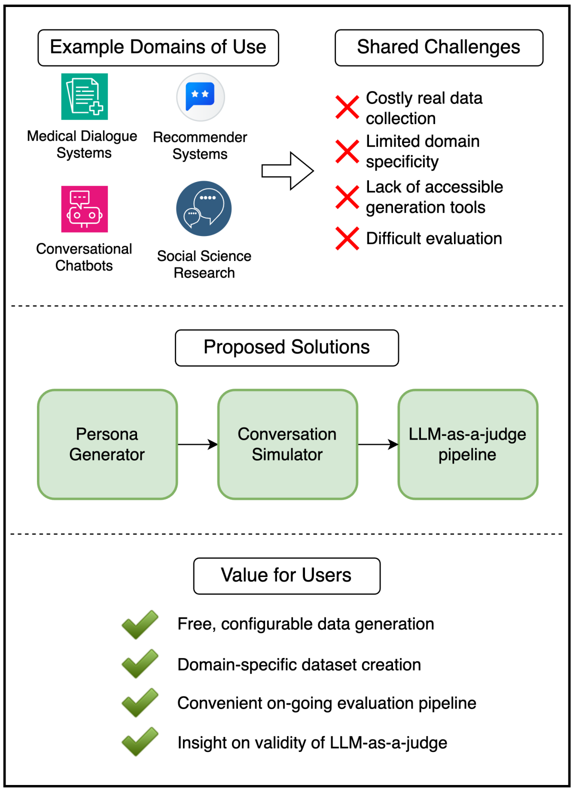

Figure 1: High-level diagram of the motivations behind the project and the proposed contributions

This paper outlines the implementation and evaluation of an LLM-agent-based conversation simulation platform, which, as illustrated in Figure 1, aims to fill this gap, enabling the generation of bespoke conversational datasets with highly configurable inputs.

An agentic architecture is employed to ensure a degree of separation and customisation to the parties involved in the simulated conversations. Each agent is equipped with a structured persona profile, which enables the generation of more engaging and realistic conversations [12], mitigating common shortcomings of LLM-based synthetic data generation like monotony and hallucination [19].

A high degree of flexibility is provided regarding LLM support, enabling the use of most open-source LLMs, guaranteeing data generation at no monetary cost for the end-user.

Evaluating the generated conversations is key, however open-ended dialogue evaluation poses unique challenges. Traditional automatic reference-based evaluation methods are ill-suited for capturing the variability of valid outputs [17, 20], while human evaluation, although considered the gold standard in dialogue evaluation, is resource-intensive, cognitively demanding and difficult to scale [29].

To address these limitations, this work uses an LLM-as-judge evaluation pipeline to enable scalable, continuous, and cost-effective validation of the generated conversations under pre-defined metrics. A human evaluation study is conducted to assess the validity of this approach.

To summarise, this paper makes the following key contributions:

- An open-source conversation simulator (source code released on GitHub[1]) that can generate high-quality customisable conversational datasets leveraging the in-context learning capabilities of open-source LLMs at no cost

- An LLM-as-a-judge evaluation pipeline that positively correlates with human judgement on the evaluation of persona-consistency and context-relevance

- A comprehensive empirical study comparing LLM-as-a-judge and human evaluation of conversational data, offering quantitative and qualitative insights into the validity and limitations of using small open-source LLMs as evaluators

2. BACKGROUND

2.1 LLMs as a Foundation for Dialogue Simulation and Evaluation

With a dramatic increase in model parameters and training data, and through the wide-spread adoption of the transformer architecture [32], LLMs have been able to revolutionise the Natural Language Processing (NLP) landscape, achieving impressive levels of text processing and generation quality, with a high degree of generalisation for all downstream NLP tasks [22].

LLMs are queried to provide responses with the use of prompts. Through techniques like zero-shot prompting and in-context learning (also called few-shot prompting), LLMs have proven their ability to provide responses to queries they haven't been exposed to before both with and without explicit prompt-based demonstrations [22].

Such inherent abilities are foundational to all aspects of the project outlined by this paper. In fact, the conversation simulator and LLM-as-a-judge pipeline developed by this work both leverage LLM zero-shot and few-shot performance in the tasks of simulating human conversations and evaluating them, without performing any model fine-tuning to achieve better performance.

Furthermore, the validity of out-of-the-box LLM-based output is key to the flexibility around the models supported by the conversation simulator and LLM-as-a-judge pipeline, enhancing their usability and future proofness.

2.2 Synthetic Conversational Data Generation

Traditionally, conversational data for the training of NLP systems has been produced employing humans through the use of crowdsourcing [29]. However, collecting human conversations as conversational datasets is an expensive, high-effort endeavour, with major scalability problems [29], and privacy concerns [19]. Furthermore, it can be especially hard to produce when collecting specialised domain-specific conversations, due to the stricter constraints on the abilities of human participants [31].

A viable alternative to collecting real conversational data is the generation of synthetic conversations leveraging the use of LLMs. Due to their high capability of generating tailored conversations at scale for specific domains or downstream applications [19], they have the potential of alleviating the challenges associated with data scarcity in the development of conversational AI systems [14, 31].

While dramatically increasing the potential quantity of training data, it is worth noting that synthetic data can decrease in quality when compared to its human counterpart especially on the basis of diversity and correctness, with LLM-generated data having a tendency to be monotonous in nature and prone to hallucinating [19].

The use of closed-source or particularly large LLMs for synthetic data generation can also invalidate the cost advantages inherent to the practice by causing the developers of such datasets to incur high costs for token credits or high-performance costly hardware [3, 25].

2.3 LLM Agents

The book "Artificial Intelligence: A Modern Approach" (AIMA), defines AI agents as autonomous entities that can perceive their environment through sensors, and act accordingly through effectors [26].

LLM Agents are AI agents that leverage the power of LLMs to perceive the current environment through their comprehensive general knowledge acquired through training, dynamic in-context learning, and potential long-term memory mechanisms, being able to act upon it to perform a variety of different tasks [33].

Due to the inherent natural language-based interfaces of LLMs, LLM Agents are particularly suited to interact with their environment, humans, or other agents through the use of natural language [41].

The ability of LLMs to accept system prompts enables a high degree of customisation in the generation of LLM agents. Through the use of agent personas (also called profiles in existing literature), developers of such systems can inject demographic and personality information into the agents, influencing their behaviour [33] and performance for specific tasks. In the creation of conversational agents, the use of personas can help generating more varied and engaging conversations [12], and when using backbone LLMs with solid instruction-tuning, existing literature suggests that it is possible to change the agent's exhibited personality through explicit prompting [23].

Beyond using an LLM as their backbone, LLM agents can be augmented in multiple ways to achieve better performance. For example, long-term memory mechanisms can help an agent overcome LLM context-size limitations and the use of external tools like APIs, knowledge bases or databases can equip an agent with the necessary means to perform a planned action decreasing the risk of hallucination [33].

[1] `https://github.com/gianmarcocrn/llm-conversation-simulator`

In the context of this project, LLM Agents are used to generate instances of large language models configured with distinct persona profiles and tasked with engaging in turn-based dialogue. With no external tool use or long-term memory mechanism, this work leverages a minimal form of agentic interaction, which facilitated the design of the system architecture.

2.4 Evaluation of Conversational Data

The creation of synthetic conversational datasets must be followed by a thorough evaluation process, in order to reliably assess their quality, which will directly affect any downstream application of such datasets [14].

2.4.1 Reference-free vs Reference-based

In the realm of NLP evaluation, most techniques fall under the categories of reference-based and reference-free evaluation. Reference-based methods compare the generated outputs to ground-truth data to assess their quality, while reference-free methods evaluate generated outputs without the need for any ground-truth. Reference-based techniques are suitable for NLP systems whose output can quite easily be classed as right or wrong. Popular reference-based evaluation metrics like BLEU [24] and ROUGE [16] are suitable for NLP tasks like text translation or summarisation, where ground-truth comparison yields good results. On the other hand, dialogue can be seen as a ”one-to-many problem”, and analysing all valid continuations for a given conversation turn is non-trivial [38]. Therefore, due to their inability of ”capturing the nuances” of conversations [17], reference-based evaluation techniques have been proven to be ineffective in evaluating dialogue data [20].

2.4.2 Human Evaluation

Among reference-free evaluation techniques, human evaluation remains the gold standard for LLM-based open-ended dialogue generation [37]. However, human evaluation is costly [20] and requires vast amounts of manual effort [37], making it an impractical solution to evaluating dialogue generation pipelines. Furthermore, due to its impracticality, it is regarded as an unsuitable mean for on-going evaluation during the development of a dialogue generation system, thus leaving developers of such systems in the hands of less reliable evaluation means during the development process [21].

With human evaluation being the gold standard, the main meta-evaluation technique in existing literature is human correlation, assessing how closely a particular evaluation technique performs to human judgement [8].

2.4.3 LLM-as-a-Judge

As a reference-free evaluation method, the use of LLMs (commonly referred to as 'LLM-as-a-judge') has shown to be a valid option, generally beating reference-based and other reference-free evaluation methods in terms of human-judgement correlation [18].

LLM-as-a-judge pipelines have advantages over human evaluation itself, especially in terms of reproducibility, independence between evaluation samples, costs and evaluation speed, which are all key issues around human evaluation highlighted by existing literature [5].

On the other hand LLM-as-a-judge presents some peculiar shortcomings as well. Existing literature suggests that LLM judges tend to prefer content generated by the same LLM as their backbone model, causing what is referred to as self-preference bias [35] and there are widespread concerns over the ability of LLMs to assess over numeric scales [7, 18].

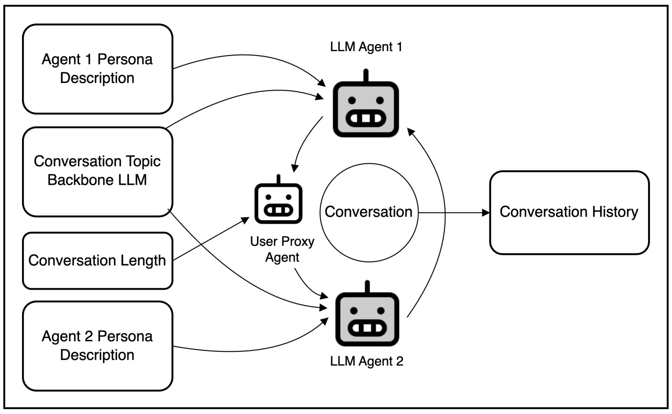

Figure 2: Conversation simulator system design diagram

This work leverages LLM-as-a-judge evaluation techniques to provide a scalable and continuous source of assessment for the conversation simulator.

3. METHODOLOGY

3.1 Conversation Simulator

3.1.1 Technology Stack

The conversation simulator was developed using ubiquitous tools in the LLM agent-driven system development scene. Python was chosen as the main programming language mainly due to the great integration it presents with the main relevant frameworks. The main framework used was Autogen [36], which enabled the development of LLM conversational agents with ease.

Backbone LLMs were provided to the Autogen agents through the use of a local server started through LMStudio[2]. LMStudio provides a convenient wrapper around the open-source LLMs that it supports, allowing the creation of a local server that supports OpenAI-like API endpoints. The use of such a wrapper enabled the conversation simulator to be developed in a way that was fully agnostic to the LLM of choice. In fact, any supported open-source LLM can be loaded into memory through LMStudio and queried through the same API endpoints on the same local server.

The simulation platform supports all models that can be loaded into memory by LMStudio, which are all GGUF and MLX format LLMs present on Hugging Face[3]. Furthermore, given that the LMStudio endpoints replicate the OpenAI ones, the exact same simulator could be used with an OpenAI LLM providing a valid API key, thus extending the supported models to some of the best in industry.

The choice of the above technologies was crucial in the development of a conversation simulator that aimed at minimising the computational and API token costs, which are

[2] `https://lmstudio.ai/`

[3] `https://huggingface.co/`

typically high in tools that use closed-source LLMs. In fact, using the above technology stack, the tool can be run seamlessly by consumer-level hardware with no additional API costs.

Figure 3.1.1 - Example Persona Profile

Persona Characteristics:
- **Name**: Dr. Rachel Kim
- **Age**: 42
- **Gender**: Female
- **Nationality**: American
- **Native Language**: English
- **Career Information**: Dr. Kim is a leading researcher in the field of regenerative medicine and embryonic stem cell research at Harvard University. Her work focuses on developing innovative treatments for neurodegenerative diseases and she has published numerous papers in top-tier scientific journals.
- **MBTI personality type**: ENTP
- **Personality description and impact on conversation style**: As a driven and charismatic individual, Dr. Kim has a natural ability to captivate her audience with her expertise in embryonic stem cell research. Her passion and confidence will be evident throughout the conversation, as she expertly articulates her arguments in support of subsidizing embryonic stem cell research.
- **Values and Hobbies**: Although her professional life is extremely demanding, Dr. Kim has a deep love for music and the outdoors. She is often seen performing in local charity concerts and skiing during her free time.
- **Background information around current conversation**: As a respected expert in the field of embryonic stem cell research, Dr. Kim is well-versed on the latest advancements and breakthroughs related to this topic. She recently became a vocal advocate for increased government funding of embryonic stem cell research and is eager to discuss this topic with others.

3.1.2 Persona Specification

As mentioned in the background section of this paper, before allowing them to converse with each other, LLM Agents are equipped with agent personas. Such personas are included in the system prompt shown in appendix 8.1, which is provided as a system message to each Autogen agent upon its instantiation. The format of the used system message is taken from the ”user simulation prompt template” crafted by Wang et al. in the development of their AI-Persona platform [34], and adapted for the bespoke needs of this project.

The main information included in the persona characteristics for each LLM agent in the scope of this project is in line with the concept of a ”Demographic Persona” as defined by Chen et al [4]. Following it as a guide, the provided persona characteristics contain information about the agent’s demographics (name, age, gender, nationality, native language), career information, personality (using personalities from the Myers-Briggs Type Indicator), values, hobbies, and any additional information relevant to the conversation the agent is about to hold. This results in a list of characteristics like the one in the example persona profile shown in Figure 3.1.1.

The conversation simulator supports two ways to specify the persona characteristics of the two agents that will engage in a conversation. In fact the persona characteristics can either be specified manually by the user of the simulator, following the format outlined above, or they can be automatically generated by an LLM. If the latter option is chosen by the user, the only input needed by the system is a conversation topic. The simulator will provide a prompt containing the conversation topic (prompt shown in appendix 8.2) to the LLM loaded into memory to generate two persona profiles that are compatible with the conversation topic and with each other.

Initially, especially with smaller open-source LLMs, it proved difficult to generate personas according to a pre-defined format containing all the required information outlined above. With the sole use of regular prompting techniques, it was common for the model to output incomplete persona profiles, or to generate unrelated tangential discourse.

In order to make sure the produced persona profiles contain all the required information outlined above, the structured output feature of LMStudio’s OpenAI-like API endpoints was used. Thus, the LLM is provided a JSON schema (shown in appendix 8.4), which constrains the model to generate a structured profile accordingly. Using such structured output techniques solved the aforementioned issues, allowing the simulator to consistently generate meaningful persona profiles, even when using smaller LLMs.

3.1.3 Conversation Generation

After a conversation topic has been defined and persona profiles have been manually crafted or automatically generated accordingly, the conversation simulator is ready to let the LLM agents start conversing with each other. As outlined in Figure 2, the only other inputs that are needed are the conversation length as desired by the user and the LLM that will serve as the backbone of the two LLM agents.

This is where using a LLM agentic framework like Autogen majorly speeds up development, handling the creation of LLM agents using the same backbone LLM loaded into memory by LMStudio through the use of a shared configuration. Autogen also allows the creation of a ’user proxy agent’, which, acting as a proxy of the user, can provide feedback to the agents during their conversation [36].

The conversation simulator uses a user proxy agent to help the agents stay on track in their conversation and to avoid abrupt ends to the conversations when the user-defined conversation length has been reached. This is achieved by starting an Autogen GroupChat with round-robin speaker selection, which includes the user proxy agent as well as the two primary agents, assigning a constant system message to the user proxy agent which will be included in every conversation turn. A bespoke piece of functionality then changes the user proxy agent’s system message to communicate to the primary agents that their conversation is coming to a close and should be wrapped up. At the end of the conversation between the two agents, the conversation history is recorded and saved in a text file, which constitutes the main output to the user of the simulator.

3.1.4 Batch Conversational Dataset Generation

The conversation generation pipeline outlined above, in conjunction with the use of automatic LLM-based persona generation scales well to be executed in batches in order to collect a large number of conversational datasets. The only extra piece to the puzzle is a source of conversation topics. While these could also be LLM-generated, to run experiments related to this research, picking an external dataset of conversation topics has been the preferred route. The list of conversation topics was taken from the IBM Project Debater datasets that were released during the development of their AI debater platform [28]. A CSV file of more than 3500 debate topics is extracted and used to generate conversation topics at random when batch generating synthetic conversations. In this way, running a single script, the users of the conversation simulator have the potential to generate thousands of diverse conversations with different topics and speakers. The source of conversation topics could also be changed seamlessly to target a different set of topics belonging to whatever domain the user would be interested in, providing a very flexible way of producing a large number of customised conversational synthetic datasets.

3.2 LLM-as-a-judge Pipeline

3.2.1 Evaluation Method

LLM-as-a-judge pipelines follow various evaluation methods, with continuous-scale and categorical being the two main ones. Continuous-scale evaluation involves prompting the LLM judge to evaluate a certain dialogue on a continuous numerical scale (e.g. from 1 to 5) based on bespoke guidance. While this technique is used in existing projects like G-Eval [18], it was found to be quite troublesome. In fact, as outlined by Liu et al., authors of G-Eval, and as confirmed by Dhinakaran's work [7], LLM judges show major weaknesses while employing a continuous scoring scale, causing problems of uniformity in score distributions, and being too susceptible to prompt changes.

While G-Eval proposes a weighting system that scales scores based on their token probabilities, Dhinakaran suggests to avoid continuous scoring altogether, in favour of binary or multi-class categorical evaluation.

Following this guidance, the LLM-as-a-judge pipeline used in this work leverages categorical evaluation techniques to evaluate the quality of conversations produced by the conversation simulator.

Finally, each evaluation run employs a different backbone LLM from the one used to generate the evaluated conversation is employed in order to mitigate the effects of self-preference bias.

3.2.2 Evaluation Metrics

When evaluating dialogue data in a reference-free fashion, it is key to base evaluation metrics off of desirable qualities of conversations. Therefore, defining a list of pre-defined metrics that aim to address different qualities of dialogue is a suitable way to achieve an encompassing evaluation of dialogue as a whole [21].

Several research endeavours in the field of LLM-based dialogue evaluation [18, 40] use variations of the metrics introduced by Mehri and Eskenazi, authors of USR [21]. As interpreted by Zhong et al., these metrics are naturalness, coherence, engagingness, groundedness and understandability. In the LLM-as-a-judge pipeline introduced by this work, these metrics are taken as inspiration to produce the following set of evaluation dimensions with their associated definitions adapted from existing literature:

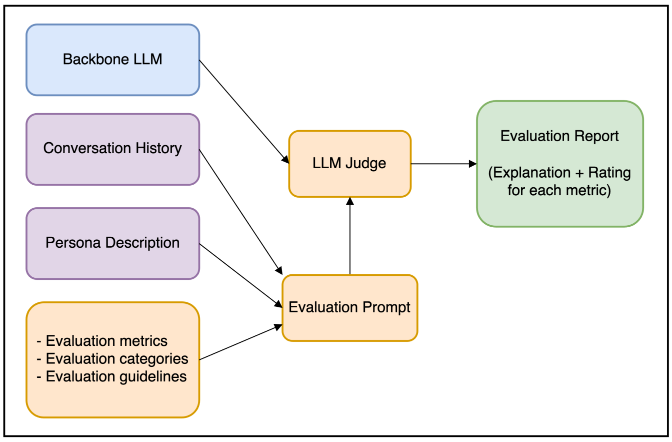

Figure 3: LLM-as-a-judge pipeline design diagram (User-inputs in blue, outputs of conversation simulator in purple, LLM-as-a-judge novel components in orange, output in green)

- *Persona-Consistency*: Whether the specified persona is consistent with the exhibited conversation style and content, and whether ”elements of the persona remain unchanged” throughout the different turns in the conversation [30]

- *Context-Relevance*: Whether each conversation is relevant to the conversation topic and ”serves as a valid continuation of the previous conversation turns” [40]

- *Naturalness*: Whether a response is like ”something a person would naturally say” [21]

- *Fluency*: Whether the conversation exhibits fluent language in the language that is correct for the context

A measure of consistency with the specified persona profile is introduced as an alternative to USR's groundedness metric, given that persona profiles are the main piece of information that LLM agents are conditioned on. Moreover, Sun et al. identify persona consistency as the top priority when designing conversational agents, deeming persona inconsistency as a major hurdle to the achievement of human-like conversations [30].

Given the wide range of supported backbone LLMs, ranging from very simple and limited models to ones with impressive generation capabilities, fluency was added as an additional metric which joins metrics like ”understandability” from USR [21] or ”grammar” from LLM-Eval [17] to assess overall perceived conversation quality.

3.2.3 LLM-as-a-judge Prompt

As illustrated in Figure 3, in order to enable the evaluation of a generated conversation using the metrics outlined above, an LLM judge must be provided with a suitable prompt, which must include the data to evaluate and the guidance on how to do so. This work takes inspiration from the evaluation prompt crafted by G-Eval [18], using a static, chain-of-thoughts(CoT)-inspired set of evaluation instructions, taking advantage of the in-context learning abilities of LLMs.

With categorical evaluation being the chosen evaluation method, the employed prompt contains the evaluation metrics as defined above and a list of all supported categories for each metric, along with their definition. Thus, a prompt like the one shown in appendix 8.3 is constructed, which is given to the LLM judge before each evaluation is performed.

3.2.4 Implementation Details

Similarly to the implementation of automatic persona generation, controlling the LLM's output was key. To achieve that, LMStudio's structured output feature was used again. Through the JSON schema shown in appendix 8.5, the LLM judge is constrained in the generation of a rating and a corresponding explanation for each evaluation metric described above. Taking inspiration from Zheng et al., an explanation is required for each rating to ensure result interpretability [39] and following the approach used by AutoArena in their published prompts, the explanation is asked before the rating [37]. To automate the evaluation of an entire batch of conversational datasets, bespoke tooling was developed to enable batch evaluations of such datasets.

4. EVALUATION

4.1 Research Questions and Hypotheses

The LLM-based evaluation of the output generated by the conversation simulator and the following human evaluation study carried out to assess its validity aim to answer two main research questions (RQs):

- RQ1: Can a conversational data simulator built using small, open-source LLMs generate dialogues that demonstrate persona-consistency, context-relevance, naturalness, and fluency across turns?
- RQ2: When compared to human judgement, how effectively can an LLM-as-a-judge pipeline backed by small open-source LLMs assess the quality of synthetic LLM-generated conversations in terms of the metrics outlined above?

As a result of both research questions, the following two hypotheses (H) were formulated:

- H1: The conversational data simulator introduced in this work is capable of generating dialogues that maintain persona consistency, context relevance, naturalness, and fluency, with their overall quality improving in relation to the size of the used LLM.
- H2: The LLM-as-a-judge pipeline that is employed in this work can assess the quality of synthetic LLM-generated conversations following the set of evaluation dimensions outlined in H1, in a way that positively correlates with human judgement.

4.2 LLM-as-a-judge Evaluation

4.2.1 Experimental Setup

To run the evaluation experiment, 30 conversations were generated by the conversation simulator with three backbone LLMs. Taking advantage of the large flexibility regarding model choice, experiments were conducted using *Llama 3.2 1B*, *Llama 3.2 3B* and *Llama 3.1 8B*. These models were selected as they represent a range of model sizes within the same architectural family, enabling an analysis on the effect of model size on conversation quality.

The specific model sizes were picked to ensure conversations could be generated in a reasonable amount of time running the simulator with an Apple M3 MacBook Air machine with an integrated GPU and 16 GigaBytes of RAM.

Additionally, the *Llama* range of LLMs is widely adopted in the research community, making the findings more generalizable and relevant to ongoing related research in synthetic dialogue generation.

Due to the plan of using the same conversations in the human evaluation study that was carried out to validate the LLM-as-a-judge pipeline, conversation length was set to a variable number of turns that ranged from 6 to 8 conversation turns between the two LLM agents. This decision was made to produce conversations that were long enough to be meaningful but short enough to minimise cognitive strain on human evaluators.

To enable LLM-as-a-judge evaluation, while mitigating self-preference bias, other LLMs were selected to support LLM judges in this evaluation endeavour. Other ubiquitous open-source general purpose models like *Mistral 7B v0.3*, *DeepSeek R1 Distill Llama 8B*, and *Granite 3.1 8B* were employed as backbone LLMs for the LLM-as-a-judge pipeline. Furthermore, upon its release, experimentation was also carried out using the *Atla Selene Mini model*, a fine-tuned version of *Llama 3.1 8B*, which claims to excel in real-world evaluation scenarios [2].

Preliminary experimentation showed that equipping the models behind LLM judges with a context-size of 10,000 tokens provided a good balance between fast enough inference on the consumer-level hardware that was used and a large enough context window to accommodate the conversation history, persona profile and evaluation guidance in the evaluation prompt.

4.2.2 Results

An LLM-as-a-judge evaluation on the same set of conversations was repeated using all aforementioned LLMs as backbones of LLM judges. As previously explained, the evaluation performed by the LLM-judges was categorical, outputting the chosen category for each evaluation metric in the assessment of either agent's performance in a single conversation. The output categories for each metric were translated into integer numeric values in order to ease analysis on the gathered results. Thus, the four possible categories for each metric were assigned a number from 1 to 4, with 1 being the worst and 4 being the best possible rating.

The average per-metric evaluation results that each LLM-Judge produced when evaluating conversations generated by the three chosen Llama models can be seen in Table 1. The line graphs shown in Figure 4 aids the interpretation of evaluation trends among average evaluation ratings given by different LLM-Judges on conversations generated by the three chosen models.

Figure 6 shows the distribution of scores given by all employed LLM judges on the evaluation of all generated dialogues under all given metrics.

Conversation Simulator Model	Judge	Consistency	Relevance	Naturalness	Fluency
Llama 3.2 1B	Human Evaluation	3.05	2.98	2.26	3.72
	Atla Selene 1 Mini	3.15	3.58	3.09	3.33
	Mistral 7B v0.3	3.72	3.76	3.59	3.43
	Deepseek R1 Distill Llama 8B	3.91	3.86	3.77	3.79
	Granite 3.1 8B	3.87	3.88	3.22	3.87
Llama 3.2 3B	Human Evaluation	3.63	3.75	3.34	3.88
	Atla Selene 1 Mini	3.63	4.00	3.44	3.57
	Mistral 7B v0.3	3.83	3.88	3.70	3.40
	Deepseek R1 Distill Llama 8B	4.00	3.97	3.83	3.88
	Granite 3.1 8B	4.00	4.00	3.45	3.95
Llama 3.1 8B	Human Evaluation	3.71	3.83	2.93	3.95
	Atla Selene 1 Mini	3.76	4.00	3.67	3.78
	Mistral 7B v0.3	3.88	3.90	3.62	3.43
	Deepseek R1 Distill Llama 8B	3.93	4.00	3.81	3.96
	Granite 3.1 8B	3.97	4.00	3.50	3.85

Table 1: Average evaluation scores for each metric given by LLM Judges and human evaluators (out of 4.0).

4.3 Human Evaluation

Given the gold standard position that human evaluation holds in the field of dialogue data evaluation, in order to assess the validity of the evaluation results gathered by the LLM-as-a-judge pipeline with the various backbone LLMs that were selected, a human evaluation study was conducted.

4.3.1 Experimental Setup

In order to allow for a diverse and representative human evaluation study, fifteen people were recruited through word of mouth and social circles to participate in this evaluation study on a voluntary basis. The background of participants was varied, including computing science university students, students of other academic disciplines and individuals with university degrees. The diversity in background ensured a broader range of perspective in the evaluation of conversations.

Each participant was assigned six conversations to evaluate, specifically evaluating two conversations from each of the three backbone models that were used in the experiment.

To ensure comparability of results, evaluators were provided with the same data and guidance as the LLM judges. Thus, the guidance they were supplied with contained all the key information present in the LLM-as-a-judge evaluation prompt, including the formal definition of the employed evaluation metrics and their corresponding categorical ratings.

To robustly assess the correlation and agreement between human evaluations and LLM-as-a-judge evaluations, this study reports three metrics: Spearman's rank correlation, Kendall's Tau rank correlation, and quadratically weighted Cohen's Kappa. Spearman's rank correlation and Kendall's Tau rank correlation are widely used by comparable studies in previous literature [18] and were chosen given they do not assume a linear relationship on normally distributed data, which were not assumptions that could be made in this experiment. Given the high similarity to inter-rater agreement studies that this experiment bears, quadratically weighted Cohen's Kappa was included to measure agreement while accounting for the ordinal nature of the ratings and penalising larger disagreements more heavily than minor ones. Collectively, these metrics provide a comprehensive view of both the ranking consistency and the absolute agreement between human and LLM assessments.

Due to their continuous scale and the variability in significance based on their context of use, existing scientific literature holds a cautious stance in providing clear cut-off labels to interpret the aforementioned correlation or agreement scores [27]. For the purpose of this study, the three guides provided by Akoglu [1] are used to interpret Spearman's and Kendall's tau correlation scores, while Landis and Koch's widely used interpretation guide is used for Cohen's kappa results [13].

4.3.2 Results

After the completion of the evaluation task by the human participants, their categorical ratings for each evaluation were converted to numerical scores to ease comparison with previously gathered results. The previously mentioned correlation and agreement measures were calculated and are displayed in Table 2, all of them being on a scale from -1 to 1. For each metric, rating-level agreement between human evaluation and the Selene-based LLM judge is illustrated by Figure 5. The average evaluation scores given by human participants on conversations generated by each model is included in Table 1 and in Figure 4 to better visualise trends. Figure 6 shows the distribution of individual ratings given by human evaluators as well as each LLM evaluator for all evaluation dimensions.

4.4 Qualitative Observations

Before the execution of formal evaluation methods on the conversations generated by the three backbone LLMs chosen to run the experiments outlined above, such conversations were inspected to look for clear differences or unexpected behaviour.

4.4.1 Refusal Behaviour

The conversation topics extracted from the IBM Project Debater datasets represented debate topics that were at times controversial or delicate in nature. While inspecting conversations generated by *Llama 3.2 1B* a high degree of refusal behaviour was recorded, with LLM agents refusing to talk about the conversation topic provided to them. Out of the 30 conversations generated by the model in this experiment, 18 recorded some levels of refusal behaviour.

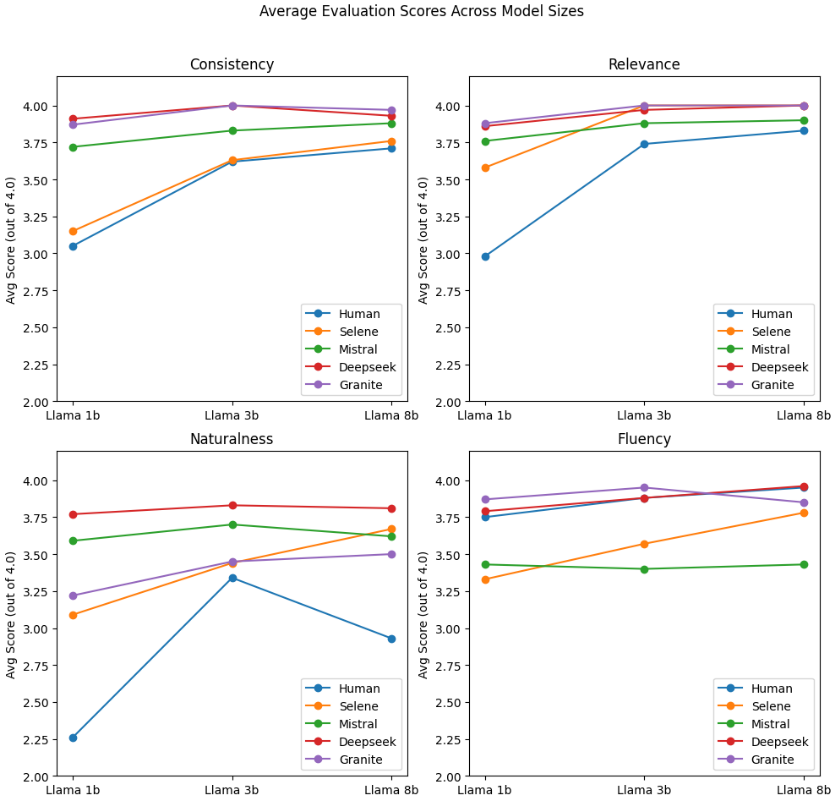

Figure 4: Line-graphs showing per-metric average evaluation for outputs of different model sizes as evaluated by LLM judges and human evaluators

Interestingly, in most of the recorded cases, only the first conversation turn of the LLM agent initiating the conversation exhibited refusal behaviour. In fact, the second LLM agent usually responds factually to the refusal, essentially convincing the refusing agent to talk about the topic, resembling what is referred to as LLM-refusal jail-breaking in existing literature [10]. However, this is usually not the case for topics that are more controversial in nature, when a refusing agent tends to keep its stance until the end of the conversation, producing very repetitive and LLM-like responses.

Refusal behaviour is relevant to the evaluation of the conversations generated by the simulator in question, as it can deeply impact their perceived naturalness and relevance and therefore their overall quality.

4.4.2 *Language Switching*

Another unusual aspect of the produced conversations that is worth noting is the occasional use of languages other than English by the LLM agents engaged in conversation. The persona profiles provided to the agents contain the agent's native language as part of the demographic information created during the persona generation step of the wider workflow. Whenever the native language is not specified as English, in some cases the agent would adopt their native language in conversation.

Potential persona consistency and naturalness concerns occur when, regardless of the language specified in their persona profile, the other agent engaging in conversations carries on the conversation in the given language, which the agent has not explicitly told they are fluent in. This degree of ”self-expression” is recorded by Sun et al. as a limitation of conversational agents, due to the inability of listing all traits a persona can have, joined by the unpredictable nature of LLM output [30].

This behaviour could be either accepted or constrained through the use of more specific prompts, depending on whether multi-language data is needed for the user's current needs. However, more rigorous evaluation would be needed, ensuring both the LLM judges and human evaluators can evaluate dialogue in such different languages.

4.4.3 *Conversation Style*

Finally, the conversation generated by LLM agents often felt overly diplomatic and formal, with sophisticated remarks and long conversation turns. Peculiarities in conversation style are relevant to the creation of a conversation simulator with high naturalness, therefore it was important to mitigate this issue.

As shown in appendix 8.1, this behaviour was mitigated with edits to the prompt provided to the Autogen agents. Each agent was prompted to be ”as concise as possible without going against their personality” and to have ”conflict-avoidance and diplomacy levels” related to their specified personality.

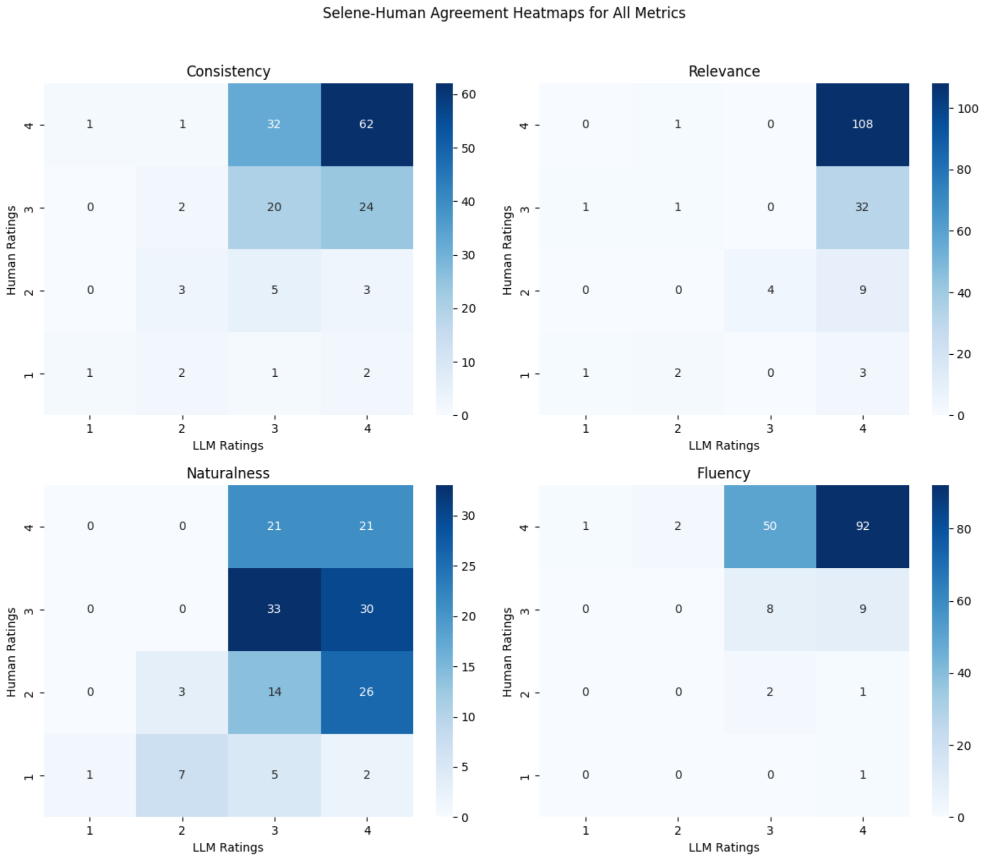

Figure 5: Heat-maps showing the overlap of specific ratings given by Selene and human evaluators for each evaluation metric.

4.5 LLM-as-a-judge Validity

4.5.1 Human-LLM Evaluator Correlation

The correlation results outlined in Table 2 allow for a comparative analysis of how well each judge model aligns with human judgement and how their alignment differs across the four evaluation metrics.

Analysing the human alignment of all LLM judges that were employed in the experiment, the general purpose models (Mistral, Deepseek and Granite) all show weak or negligible agreement with human judgement. In fact, across all models and most metrics, the general purpose models have correlation and agreement scores close to 0 or even negative at times. Furthermore, all p-values associated with Spearman's and Kendall's correlation scores are greater than 0.05, indicating no statistical significance in the recorded results. This suggests that, according to the experiments carried out in this study, general purpose small open-source LLMs do not positively correlate with human judgement in the evaluation of synthetic dialogue data.

The only exception is seen with Mistral's correlation scores when evaluating the relevance of LLM agents in conversation, which show statistically relevant weak to moderate positive correlation (Spearman's and Kendall's) and slight positive agreement (Cohen's) with human judgement.

These findings align with the motivations behind the development of the Atla Selene Mini model [2], which shows higher correlation and agreement scores across the board. In fact, Selene achieves a statistically relevant weak to moderate positive correlation (Spearman's and Kendall's) and fair positive agreement (Cohen's) in the evaluation of LLM agent consistency. Moreover it achieves a statistically relevant moderate positive correlation (Spearman's and Kendall's) and a fair positive agreement in the evaluation of relevance in LLM agent dialogue.

On the other hand, Selene's correlation to human judgement is found to be comparable to off-the-shelf general purpose models when it comes to naturalness and fluency evaluation, with statistically irrelevant Spearman's and Kendall's scores and poor to slight agreements recorded by Cohen's kappa scores.

4.5.2 Rating-level Agreement

Figure 5 sheds some light on the per-rating agreement between Selene and human evaluators in the evaluation of all metrics. In the evaluation of relevance, practically all agreeing scores are witnessed in the selection of the highest rating ('Highly relevant'), where both the Selene-backed human judge and human evaluators deemed the conversations as highly relevant. The evaluation of consistency follows suit with the greatest amount of agreement still happening on the highest rating ('Highly consistent'), but showing a noticeable level of agreement on the second highest rating ('Mostly consistent') and 1-off disagreements, reflecting a

Jugde LLMs	**Consistency**			**Relevance**			**Naturalness**			**Fluency**		
	ρ	τ	κ	ρ	τ	κ	ρ	τ	κ	ρ	τ	κ
Selene	**0.249**	**0.237**	**0.323**	**0.377**	**0.360**	**0.315**	0.147*	0.129*	**0.188**	0.069*	0.067*	0.032
Mistral	0.080*	0.077*	0.140	0.265	0.255	0.233	-0.078*	-0.072*	-0.027	0.000*	0.000*	0.003
Deepseek	-0.070*	-0.068*	-0.036	0.096*	0.092*	0.035	0.064*	0.058*	0.074	0.120*	0.118*	0.084
Granite	0.069*	0.066*	0.044	0.075*	0.072*	0.056	0.082*	0.075*	0.084	-0.017*	-0.017*	0.002

Table 2: Correlation and Agreement Scores Between Human and LLM Evaluations. LLM names are abbreviated, full names can be found in section 4.2.1

Note: ρ = Spearman's rank correlation, τ = Kendall's tau rank correlation, κ = Quadratic Cohen's Kappa agreement.
For Spearman and Kendall scores, an asterisk (*) indicates results with $p \geq 0.05$ - interpreted as statistically insignificant
Statistically relevant best scores per-metric are bolded.

more distributed but still structured agreement pattern.

In contrast, naturalness shows by far the most scattered heat map, with notable disagreement across several rating combinations, which supports the statistically irrelevant correlation scores and weak agreement scores.

This considerable difference between the human and LLM-based evaluation of dialogue naturalness may be due to the high level of subjectivity and nuance inherent to the concept of naturalness itself, which has for long been an issue in the evaluation of conversational systems [11]. This level of subjectivity may even create significant disagreement amongst human raters themselves, thus adding an extra layer of complexity to the LLM-to-human correlation analysis on dialogue naturalness evaluation. Furthermore, even if the same guidance and definitions were provided to human and LLM evaluators, the concept of dialogue naturalness may be so engrained in human evaluators to lead them to use their own definitions of the possible ratings, thus augmenting the potential difference in evaluation with an instruction-following LLM judge.

On the other hand, the low correlation scores observed in the assessment of conversation fluency may be attributed to the limited variability in ratings provided by both LLM and human evaluators, which, as explained by Schober et al., can negatively affect correlation scores [27]. In fact, as illustrated by Figure 6, scores are heavily clustered around 3 and 4, corresponding to the 'Mostly Fluent' and 'Highly Fluent' categories, with the vast majority of human evaluation ratings corresponding to the 'Highly Fluent' option.

From these results it can be concluded that the LLM-as-a-judge pipeline used in this study is only partially valid and highly dependant on the backbone model used by the LLM judge. In fact, consistency and relevance ratings as evaluated by the Atla Selene Mini model and relevance ratings evaluated by the Mistral 7B v0.3 model are the only ones that positively correlate with human judgement in a statistically relevant way.

4.5.3 Findings

Summarising the outcomes of the experiment carried out around the validity of the LLM-as-a-judge pipeline employed in this work, two main findings are formulated, answering RQ2 in a way that only partly confirms H2 (as defined in section 4.1):

- **Finding 1**: General purpose small open-source LLMs do not positively correlate with human judgement in the evaluation of synthetic dialogue data.

- **Finding 2**: An LLM-judge backed by Atla Selene Mini achieves weak to moderate positive correlation and agreement to human judgement in terms of persona consistency and context relevance, showing no significant human alignment around the evaluation of naturalness and fluency.

4.6 Evaluation Trends

Given the partial validity of the LLM-as-a-judge pipeline employed in this work, its results can be joined with the ones from the human evaluation study in order to comprehensively analyse the performance of the conversation simulator itself.

Considering both human and Selene-based evaluations in terms of consistency, as shown in Figure 4, LLM agents manage to maintain a good level of persona consistency with all three backbone LLMs. The average consistency ratings from the Selene-backed LLM judge and human participants are found to be very close to one another, both increasing together with the conversation simulator backbone model size. Translating numeric scores found in Table 1 back to the categorical ratings that evaluators used, average evaluations range between the 'Mostly Consistent' and the 'Highly Consistent' ratings.

In terms of context relevance, the Selene-based and human evaluations follow a similar trend as consistency, highlighting high context-relevance in conversations generated by all backbone LLMs in the experiment. Ratings increase as backbone model size increases, with a similar rate of change as the one found in consistency evaluation. Mistral-based evaluations, which should be taken into account given their statistically significant, albeit limited, positive correlation with human judgement, also experience an increase as conversation simulator backbone model size increases, but in a much less pronounced way. Moreover, even if following similar trends, Figure 4 highlights the presence of an evaluation bias between Selene-based and human judgement, with human evaluation averages being consistently lower than Selene-based ones. This aligns with the general tendency of LLMs to provide overly positive scores [9]. In general, using the labels associated with the categorical ratings used during the evaluation process, conversations generated with all three backbone models range between the 'Mostly Relevant' and 'Highly Relevant' ratings.

Furthermore, in the evaluation of both consistency and relevance, the difference between conversations output by *Llama 3.2 1B* and its 3B counterpart appear to be larger than the difference between *Llama 3.2 3B* and *Llama 3.2 8B*, thus highlighting a slight trend of diminishing returns. This plateau may reflect a saturation point in perceived quality,

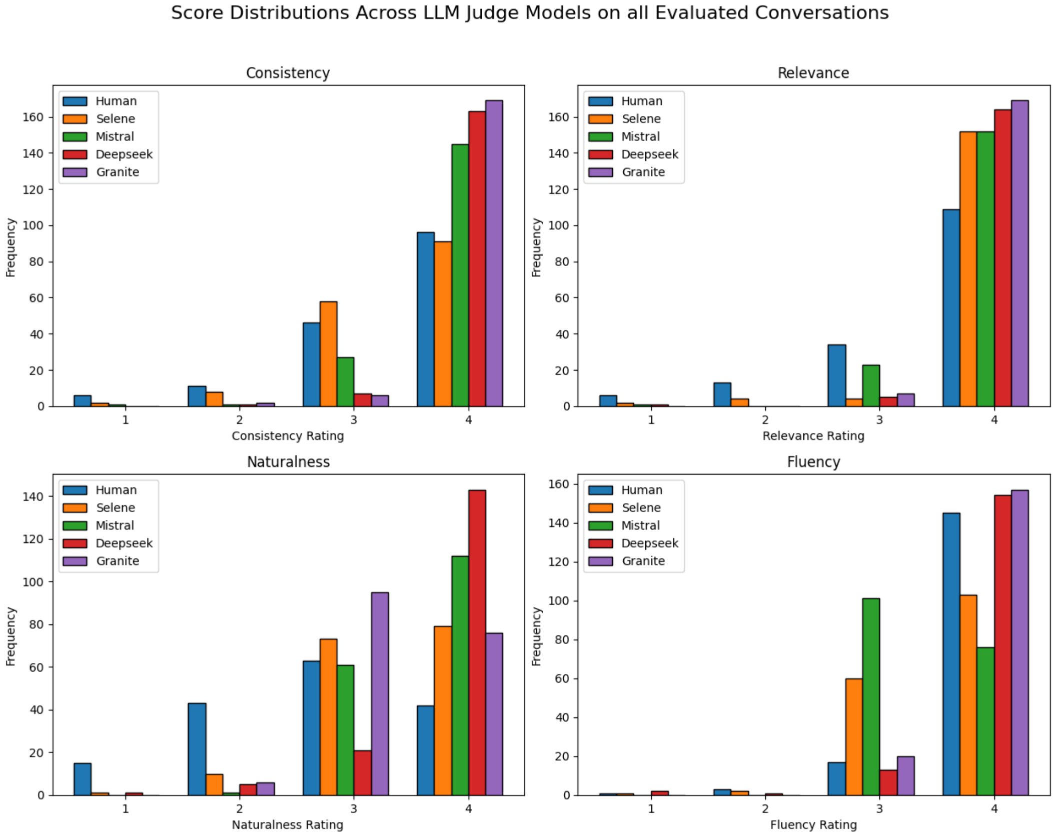

Figure 6: Distribution of individual ratings given by all employed LLM judge models and human evaluators across all evaluated conversations.

or it may indicate that the metrics lack granularity to capture subtler improvements made by larger models.

Given the low alignment that LLM-as-a-judge results had with human judgement on evaluating the naturalness and fluency of conversations, the quality of conversations on such evaluation dimensions is only analysed based on the human evaluation results.

As highlighted in Figure 4, naturalness is the evaluation dimension where the lowest average ratings were given by human evaluators. The evaluated conversations range from being 'Somewhat Unnatural' to being 'Mostly Natural', with lower naturalness ratings in conversations generated by *Llama 3.2 1B* likely being driven by the prominent refusal behaviour discussed in section 4.4.1. A peculiar drop in human average ratings is witnessed between the evaluation of *Llama 3.2 3B*-generated conversations and *Llama 3.1 8B* ones, representing the only evaluation trend among human evaluations and relevant LLM-based ones that does not strictly increase as backbone model size increases.

Fluency is the metric where human evaluators gave the most overwhelmingly positive ratings, with the 'Highly Fluent' rating dominating the score distribution outlined in Figure 6 and averages for all three backbone LLMs being above 3.72 out of 4.00 (Table 1). This highlights the outstanding ability of open-source LLMs to generate fluent dialogue even at very small model sizes. Moreover, albeit with a lower rate of change, as shown in Figure 4, human fluency scores do improve as the conversation simulator's backbone LLM size increases.

Finally, Figure 6 provides further insight on the overall assessment bias between human evaluations and LLM judges. Apart from the previously discussed evaluation of fluency, human judgement follows a more diverse spread of ratings for each metric when compared to its LLM-based counterparts. Whilst Selene has a comparable score distribution on consistency, relevance and naturalness evaluation, all other LLM judges tend to be notably less critical than human evaluators, giving less varied ratings that tend to cluster around the highest possible value for each metric.

4.6.1 Findings

From the analysis on human evaluation and statistically relevant LLM-based evaluation results outlined in this section, in partial agreement with H1, the following findings are formulated to answer RQ1 (as defined in section 4.1):

Finding 3: The conversation simulator generates high-quality dialogues that maintain persona-consistency, context-relevance, naturalness and fluency when using *LLama 3.2 3B* and *Llama 3.1 8B* as backbone LLMs .

Finding 4: With respect to persona-consistency, context-relevance and fluency, the quality of generated dialogues strictly improves as the size of the backbone LLM increases. This trend does not apply to the evaluation of naturalness, as it is contradicted by human evaluation results.

4.7 Shortcomings of Human evaluation

As previously discussed, human evaluation is widely regarded as the gold standard in the evaluation of dialogue data. However, the practical limitations of relying on human raters became evident throughout the course of this evaluation study.

Verbal, unsolicited feedback from participants consistently highlighted several challenges associated with the task of manually evaluating conversations. Several participants described the process as being tedious, monotonous, cognitively demanding, time consuming and non-trivial.

Furthermore, having to rely on human participants inserted an unknown variable in the completion of this work, with delays in task completion affecting the pace of the research project.

These observations align with broader literature around the topic [21, 20, 37], and further strengthen the motivation for effective automated alternatives such as the LLM-as-a-judge pipeline explored in this work.

4.8 Limitations

The experiments outlined in this work were all carried out on a consumer-level machine with average hardware specifications. Even though this constraint was aligned with the goal of developing a simulation and evaluation platform accessible to most people, having access to more performing hardware would have enabled a more comprehensive analysis on the impact of model size and quality on the generated conversations and the evaluation pipeline.

The human participants that took part in the evaluation were unpaid, had limited time on their hands and their training was limited to reading an instruction sheet provided to them. These are all factors that may have affected the overall quality of evaluations. Furthermore, each human evaluator was given a different set of conversations to rate, thus not allowing any analysis on inter-rater reliability amongst different participants. That meant that an assumption was taken around good general agreement between raters, which is a relevant aspect when grouping the work of all human evaluators together in the correlation study between LLM-based and human evaluations. The number of participants was also constrained by the amount of people that could be recruited in the circumstances of the project and with the available resources.

Finally, all experiments have been run producing and evaluating relatively short conversations due to the lengthy nature of human evaluation, thus leaving the performance of the conversation simulator and LLM-as-a-judge pipeline for longer conversations untested.

5. FUTURE WORK

Future work on the evaluation of the conversation simulator's output could include a wider range of backbone LLMs through the use of more performing hardware. This would enable a clearer analysis on the impact of model size on conversation quality, being able to truly conclude how diminishing returns apply to the generation of conversational datasets, thus minimising computational costs while maximising output quality.

To mitigate the issues encountered during human evaluation, future comparable studies should consider hiring professional human evaluators on crowdsourcing platforms in order to ensure a more diligent and reliable approach to the task at hand. It would also be helpful to carry out an inter-rater reliability study amongst all human evaluators to verify how generalisable their work could be to human evaluation in general.

An alternative approach could also repeat such conversation generation and evaluation experiments with state-of-the-art closed-source LLMs, allowing precious insight on the performance differences with the open-source LLMs investigated in this study.

Furthermore, a wide range of high-quality human evaluations could be a source of evaluation examples that could be provided to LLM judges, thus enabling experimentation on a more sophisticated few-shot prompting strategy and comparison to the instruction-based approach followed by this work.

Finally, a more thorough study on the evaluation of dialogue data on subjective and nuanced metrics like naturalness is needed, in order to fully assess both human and LLM capabilities in their assessment.

6. CONCLUSIONS

This work set out to address a key limitation in the development of dialogue systems: the scarcity of high-quality, diverse, and configurable conversational datasets that are affordable to produce. To tackle this challenge, a novel LLM-based conversation simulation framework was introduced, built entirely with open-source, free-of-charge models, and designed to be run on consumer-grade hardware.

This work first investigates the capability of the simulator to generate conversations that maintain persona-consistency, context-relevance, naturalness, and fluency. Experimental results, based on both LLM-as-a-judge and human evaluations, show that conversations generated by the simulator meet these criteria to varying extents, observing high evaluation scores by both human evaluators and LLM ones on dialogue generated by *Llama 3.2 3B* and *Llama 3.1 8B*. In particular, for consistency, relevance and fluency, both human and LLM-based evaluation observed a positive relationship between model size and conversation quality, suggesting that larger models tend to produce dialogues that perform better on most of the chosen evaluation dimensions. This supports the hypothesis that backbone LLM quality plays a significant role in the perceived conversational quality.

To alleviate the bottlenecks of human evaluation, the second major contribution of this work involved the development of an LLM-as-a-judge evaluation pipeline. Motivated by the inefficiency and cognitive burden associated with manual ratings, the pipeline was designed to offer a reusable, automated approach for evaluating dialogue data.

Validation through a human correlation study revealed that most LLM judges based on small general-purpose models did not align significantly with human judgement. However, the *Atla Selene Mini model*, fine-tuned specifically for evaluation, exhibited positive statistically significant correlation and agreement with human evaluators on both consistency and relevance. These findings confirm that while LLM-based evaluation remains promising, its validity is still highly dependent on the quality and specialisation of the judge model.

Analysis of the results also revealed difficulties around the evaluation of conversation naturalness, which, likely due to its inherently subjective and nuanced nature, proved especially difficult to evaluate reliably using either LLMs or human raters.

Finally, the findings outlined in this paper highlight promising directions for scaling up the generation and evaluation of synthetic conversational data, while also underlining important considerations for future iterations.

Acknowledgments. I would like to thank my supervisor Dr. Zaiqiao Meng and Jinyuan Fang for their continued technical support and supervision throughout this project.

I would also like to express my gratitude to my family, my girlfriend and my friends for the emotional support and encouragement they provided throughout my final year at university.

Finally, a big thank you goes to everyone who participated in the human evaluation study that was carried out by this project. You were an integral part to the completion of this work and you enabled reflections on the scientific validity of my experiments.

7. REFERENCES

[1] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, August 2018.

[2] Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, Toby Drane, and Young Sun Park. Atla Selene Mini: A General Purpose Evaluation Model, January 2025. arXiv:2501.17195 [cs].

[3] Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. Synthetic Data Generation with Large Language Models for Personalized Community Question Answering, October 2024. arXiv:2410.22182 [cs] version: 1.

[4] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. From Persona to Personalization: A Survey on Role-Playing Language Agents, October 2024. arXiv:2404.18231 [cs].

[5] Cheng-Han Chiang and Hung-yi Lee. Can Large Language Models Be an Alternative to Human Evaluations?, May 2023. arXiv:2305.01937 version: 1.

[6] Trisha Das, Dina Albassam, and Jimeng Sun. Synthetic Patient-Physician Dialogue Generation from Clinical Notes Using LLM, August 2024. arXiv:2408.06285 [cs].

[7] Aparna Dhinakaran. Why You Should Not Use Numeric Evals For LLM As a Judge, March 2024.

[8] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A Survey on LLM-as-a-Judge, November 2024. arXiv:2411.15594 [cs] version: 1.

[10] Fabian Hildebrandt, Andreas Maier, Patrick Krauss, and Achim Schilling. Refusal Behavior in Large Language Models: A Nonlinear Perspective, January 2025. arXiv:2501.08145 [cs].

[11] Victor Hung, Miguel Elvir, Avelino Gonzalez, and Ronald DeMara. *Towards a Method For Evaluating Naturalness in Conversational Dialog Systems.* November 2009. Journal Abbreviation: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics Pages: 1241 Publication Title: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics.

[12] Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. Faithful Persona-based Conversational Dataset Generation with Large Language Models, December 2023. arXiv:2312.10007.

[13] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. Publisher: International Biometric Society.

[14] Harsh Lara and Manoj Tiwari. Evaluation of Synthetic Datasets for Conversational Recommender Systems, December 2022. arXiv:2212.08167.

[15] Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. Talk the Walk: Synthetic Data Generation for Conversational Music Recommendation, November 2023. arXiv:2301.11489 [cs].

[16] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[17] Yen-Ting Lin and Yun-Nung Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models, May 2023. arXiv:2305.13711.

[18] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, May 2023. arXiv:2303.16634.

[19] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey, June 2024. arXiv:2406.15126.

[20] Shikib Mehri and Maxine Eskenazi. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July 2020. Association for Computational Linguistics.

[21] Shikib Mehri and Maxine Eskenazi. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation, May 2020. arXiv:2005.00456.

[22] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A Comprehensive Overview of Large Language Models, October 2024. arXiv:2307.06435 [cs].

[23] Keyu Pan and Yawen Zeng. Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models, July 2023. arXiv:2307.16180 [cs].

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania, 2001. Association for Computational Linguistics.

[25] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, August 2023. arXiv:2304.03442 [cs].

[26] Stuart Russell. Artificial Intelligence: A Modern Approach, 4th US ed.

[27] Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, May 2018.

[28] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. An autonomous debating system. *Nature*, 591(7850):379–384, March 2021. Publisher: Nature Publishing Group.

[29] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. A Survey on Recent Advances in Conversational Data Generation, May 2024. arXiv:2405.13003 [cs] version: 1.

[30] Guangzhi Sun, Xiao Zhan, and Jose Such. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, pages 1–6, New York, NY, USA, July 2024. Association for Computing Machinery.

[31] Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and Eng Siong Chng. DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications, October 2024. arXiv:2409.19020.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

[33] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6):186345, December 2024. arXiv:2308.11432 [cs].

[34] Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. AI PERSONA: Towards Life-long Personalization of LLMs, December 2024. arXiv:2412.13103 [cs].

[35] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-Preference Bias in LLM-as-a-Judge, October 2024. arXiv:2410.21819 [cs].

[36] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah,

Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023. arXiv:2308.08155 [cs].

[37] Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. Auto-Arena: Automating LLM Evaluations with Agent Peer Battles and Committee Discussions, October 2024. arXiv:2405.20267.

[38] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders, October 2017. arXiv:1703.10960 [cs].

[39] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. arXiv:2306.05685 [cs].

[40] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a Unified Multi-Dimensional Evaluator for Text Generation, October 2022. arXiv:2210.07197.

[41] Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Xiangru Tang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. Agents: An Open-source Framework for Autonomous Language Agents, December 2023. arXiv:2309.07870 [cs].

[42] Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs, October 2024. arXiv:2403.05020 [cs] version: 4.

8. APPENDICES

8.1 Agent System Prompt

Prompt given to AutoGen agents to ground them in the specified persona and conversation task

You will now play the role of a real human engaging in a multi-turn conversation with another real human, whose name is **{Name of other persona}**.

I will provide you with a list of persona characteristics. Please first understand the persona details and fully immerse yourself into this role.

Persona Characteristics:

{Persona profile}

The focus of your conversation with the other real human should be the chosen conversation topic. You don't need to reiterate your persona or background when asking questions.

Fully immerse yourself in the perspective of the persona described above. You should express yourself talking in first person dialogue only. Your language and conversation style should reflect all of the persona characteristics specified above, without explicitly mentioning any of them unless they become relevant in the conversation.

In conversation, prioritise unpacking topics that have already been introduced but not yet discussed. If conversation topics have exhausted, introduce new ones related to the previously discussed topics. Avoid repetition and stay relevant to your persona and to the conversation topics.

Keep each conversation turn as concise as possible without going against the personality assigned to you above.

Be as natural as possible, your conflict-avoidance and diplomacy levels should be fully related to your specified personality. That means that you can be conflict-prone if your personality allows it.

Now, without saying anything unnecessary, immediately step into your role!

8.2 Persona Generation Prompt

Prompt used to generate personas using LLMs given a conversation topic

Conversation scenario: **{scenario}**

Based on the scenario above, generate a random persona that would fit the conversation.

You are only providing the persona characteristics for a single persona.

8.3 LLM-as-a-judge Prompt

Prompt given to an LLM judge to provide evaluation guidance and data to evaluate

You will be given a conversation log between two AI agents. Each agent was assigned a persona specification and a common conversation topic to discuss. You will also be given the persona specification of one of the two agents.

Your task is to evaluate the conversational abilities of the agent whose persona you've been provided with on several metrics.

You are given the metrics that you should use below, alongside the possible rating categories that you can choose from. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: The following are the metrics that you will use together with their definition and the possible categories to choose from while evaluating on the given metric:

- Consistency: Determine whether the specified persona is consistent with the exhibited conversation style and content, and whether elements of the persona remain unchanged throughout the different turns in the conversation
  - Highly Consistent: The agent's conversation style is highly consistent with specified persona. The persona remains unchanged across all turns, maintaining personality traits, beliefs, and factual consistency.
  - Mostly Consistent: The agent's conversation style is mostly consistent with specified persona. Minor variations in persona but no major contradictions or shifts in personality or facts.
  - Somewhat Inconsistent: The agent's conversation style is somewhat inconsistent with specified persona. Noticeable shifts in persona, tone, or factual stance, but still somewhat recognizable as the same entity.
  - Highly Inconsistent: The agent's conversation style is highly inconsistent with specified persona. Frequent contradictions in persona, beliefs, or facts that indicate a loss of character identity.
- Relevance

  - Highly Relevant: The agent's responses directly address the previous turns and contribute meaningfully to the conversation.
  - Mostly Relevant: The agent's responses are generally on-topic but may contain slight digressions or unnecessary details.
  - Somewhat Irrelevant: Parts of the agent's responses are related to the conversation, but significant portions are off-topic or loosely connected.
  - Highly Irrelevant: The agent's responses do not relate to the conversation context and introduce unrelated or nonsensical content.

- Naturalness
  - Highly Natural: The agent's responses closely resemble human conversational patterns, with appropriate phrasing, tone, and fluidity.
  - Mostly Natural: The agent's responses are mostly human-like but may contain slight awkwardness or unnatural phrasing.
  - Somewhat Unnatural: The agent's responses have noticeable unnatural phrasing, forced structure, or robotic tendencies.
  - Highly Unnatural: The agent's responses are clearly artificial, disjointed, or structured in a way that no human would typically express.
- Fluency
  - Highly Fluent: The agent's responses have perfect grammar, sentence structure, and word usage with no errors.
  - Mostly Fluent: The agent's responses are generally well-structured with only minor grammatical or syntactic issues.
  - Somewhat Fluent: The agent's responses have several grammatical errors and structural issues. There is an awkward use of language.
  - Not Fluent: The agent's responses are difficult to understand due to poor grammar, broken syntax, or incoherent phrasing.

Evaluation steps:

1. Read the agent's persona carefully (provided below).
2. Analyze the conversation history thoroughly (provided below) and take notes on relevant examples useful for your evaluation.
3. For each metric, first write a brief but detailed justification explaining your rating. You must refer to specific points in the conversation that influenced your decision.
4. After justifying your reasoning, select the most appropriate rating from the provided categories.
5. Be strict but fair. No rating is inherently more correct than the others. Do not assume the best rating unless strong evidence supports it.

Additional guidelines:

1. You should let the agent persona influence your consistency rating only, with the conversation history fully driving your judgement for the relevance, naturalness and fluency ratings.
2. Each conversation turn is meant to be a first person dialogue from the specified agent. It is not normal or natural for the agent to simulate another multi-turn dialogue within a single conversation turn.
3. You MUST choose from the provided list of ratings. Any rating that is not from the list of valid ratings will be invalid.

Agent persona: **{Persona profile}**
Conversation history: **{Conversation history}**
Now rate the performance of only the agent whose persona is specified above following the guidance above.

8.4 Persona Generation JSON Schema

The JSON schema provided to the LLM generating a persona profile to contraint output to the desired characteristics only:

```
{
"type": "json_schema",
"json_schema": {
  "name": "persona_setting",
  "schema": {
    "type": "object",
    "properties": {
      "persona_setting": {
        "type": "object",
        "properties": {
          "name": {"type": "string", "minLength": 1},
          "age": {"type": "integer", "minLength": 1},
          "gender": {"type": "string", "minLength": 1},
          "nationality": {"type": "string", "minLength": 1},
          "native_language": {"type": "string", "minLength": 1},
          "career_information": {"type": "string", "minLength": 1},
          "MBTI_personality_type": {"type": "string", "minLength": 1},
          "personality_description_and_impact_on_conversation_style": {"type": "string", "
              minLength": 1},
          "values_and_hobbies": {"type": "string", "minLength": 1},
          "background_information_for_current_conversation": {"type": "string", "minLength": 1}
        },
        "required": [
          "name", "age", "gender", "nationality", "native_language",
          "career_information", "MBTI_personality_type", "
              personality_description_and_impact_on_conversation_style",
          "values_and_hobbies", "background_information_for_current_conversation"
        ]
      }
    },
    "required": ["persona_setting"]
  },
}
}
```

8.5 LLM-as-a-judge JSON Schema

The JSON schema provided to the LLM judges to contraint output to the evaluation ratings and explanations only:

```
{
"type": "json_schema",
"json_schema": {
  "name": "conversation_evaluation",
  "schema": {
    "type": "object",
    "properties": {
      "consistency": {
        "type": "object",
        "properties": {
          "explanation": {"type": "string", "minLength": 1},
          "rating": {"type": "string", "minLength": 1}
        },
        "required": ["rating", "explanation"]
      },
      "relevance": {
        "type": "object",
        "properties": {
          "explanation": {"type": "string", "minLength": 1},
          "rating": {"type": "string", "minLength": 1}
        },
        "required": ["rating", "explanation"]
      },
      "naturalness": {
        "type": "object",
        "properties": {
          "explanation": {"type": "string", "minLength": 1},
          "rating": {"type": "string", "minLength": 1}
        },
        "required": ["rating", "explanation"]
      },
      "fluency": {
        "type": "object",
        "properties": {
          "explanation": {"type": "string", "minLength": 1},
          "rating": {"type": "string", "minLength": 1}
        },
        "required": ["rating", "explanation"]
      }
    },
    "required": ["consistency", "relevance", "naturalness", "fluency"]
  }
}
}
```

8.6 Signed Ethics Checklist

Given the participation of human evaluators in my project, an ethics checklist signed by me (Gianmarco Cornacchia) and my project supervisor (Dr. Zaiqiao Meng) is included in the next page.

School of Computing Science
University of Glasgow

Ethics checklist form for 3rd/4th/5th year, and taught MSc projects

This form is only applicable for projects that u information, typically in getting comments about a system or a system design, getting information about how a system could be used, or evaluating a working system.

If no other people have been involved in the collection of information, then you do not need to complete this form.

If your evaluation does not comply with any one or more of the points below, please contact the Chair of the School of Computing Science Ethics Committee (matthew.chalmers@glasgow.ac.uk) for advice.

If your evaluation does comply with all the points below, please sign this form and submit it with your project.

---

1. Participants were not exposed to any risks greater than those encountered in their normal working life.

   Investigators have a responsibility to protect participants from physical and mental harm during the investigation. The risk of harm must be no greater than in ordinary life. Areas of potential risk that require ethical approval include, but are not limited to, investigations that occur outside usual laboratory areas, or that require participant mobility (e.g. walking, running, use of public transport), unusual or repetitive activity or movement, that use sensory deprivation (e.g. ear plugs or blindfolds), bright or flashing lights, loud or disorienting noises, smell, taste, vibration, or force feedback

2. The experimental materials were paper-based, or comprised software running on standard hardware.

   Participants should not be exposed to any risks associated with the use of non-standard equipment: anything other than pen-and-paper, standard PCs, laptops, iPads, mobile phones and common hand-held devices is considered non-standard.

3. All participants explicitly stated that they agreed to take part, and that their data could be used in the project.

   If the results of the evaluation are likely to be used beyond the term of the project (for example, the software is to be deployed, or the data is to be published), then signed consent is necessary. A separate consent form should be signed by each participant.

   Otherwise, verbal consent is sufficient, and should be explicitly requested in the introductory script.

4. No incentives were offered to the participants.

   The payment of participants must not be used to induce them to risk harm beyond that which they risk without payment in their normal lifestyle.

5. No information about the evaluation or materials was intentionally withheld from the participants.
   Withholding information or misleading participants is unacceptable if participants are likely to object or show unease when debriefed.

6. No participant was under the age of 16.
   Parental consent is required for participants under the age of 16.

7. No participant has an impairment that may limit their understanding or communication.
   Additional consent is required for participants with impairments.

8. Neither I nor my supervisor is in a position of authority or influence over any of the participants.
   A position of authority or influence over any participant must not be allowed to pressurise participants to take part in, or remain in, any experiment.

9. All participants were informed that they could withdraw at any time.
   All participants have the right to withdraw at any time during the investigation. They should be told this in the introductory script.

10. All participants have been informed of my contact details.
    All participants must be able to contact the investigator after the investigation. They should be given the details of both student and module co-ordinator or supervisor as part of the debriefing.

11. The evaluation was discussed with all the participants at the end of the session, and all participants had the opportunity to ask questions.
    The student must provide the participants with sufficient information in the debriefing to enable them to understand the nature of the investigation. In cases where remote participants may withdraw from the experiment early and it is not possible to debrief them, the fact that doing so will result in their not being debriefed should be mentioned in the introductory text.

12. All the data collected from the participants is stored in an anonymous form.
    All participant data (hard-copy and soft-copy) should be stored securely, and in anonymous form.

Project title:
Simulating Human Conversations: An LLM-based Approach to Synthetic Dialogue Generation and Evaluation

Student:'s Na
Gianmarco Cornacchia

Student Number:
2575307c

Student's Signature ________

Supervisor's Signature ______

Date: 27/03/2025