Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Technologies (FEM)

Bachelor Thesis

Context Consistency In AI Dialogue Systems

Evgeny Kren

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Evgeny Kren

Informatics

Thesis title

Context Consistency In AI Dialogue Systems

Objectives of thesis

The main objective of this work is to evaluate and compare selected AI dialogue systems in terms of maintaining context using experimental testing scenarios.
The partial goals of this work include:
- Study and analyze relevant literature and information sources on AI dialogue systems and AI contextual consistency.
- Evaluate and compare different approaches to maintaining context in AI dialogue systems.
- Develop different testing scenarios for the AI dialogue systems with a focus on ascertaining the context consistency of selected chat-bot solutions.

Methodology

The theoretical part of the work is based on the study and analysis of professional and scientific information sources. The thesis addresses AI context consistency with a specific focus on maintaining context in AI dialogue systems. The different approaches to maintaining context in AI dialogue systems will be analyzed and compared based on chosen criteria. An experimental testing conversations for the AI dialogue systems will be used to evaluate selected solutions and their ability to maintain context consistency. Based on the synthesis of knowledge from the theoretical part and the evaluation of the results of the practical part, the conclusions of the work will be formulated.

The proposed extent of the thesis

40-50

Keywords

AI, Dialogue systems, Natural language processing, Contextual responses, Machine learning, Chat-bot, Personal assistant, User computer interaction

---

Recommended information sources

Daniel Jurafsky, James Martin. Speech and Language Processing, 2nd Edition. Prentice Hall: 2008. ISBN-13: 978-0131873216

Michael McTear, Zoraida Callejas, David Griol. The Conversational Interface: Talking to Smart Devices. Springer: 2016. ISBN-13: 978-3319329659

Peter Norvig, Stuart Russell. Artificial Intelligence: A Modern Approach, Global Edition 4th Edition. Pearson: 2021. ISBN-13: 978-1292401133

Sumit Raj. Building Chatbots with Python: Using Natural Language Processing and Machine Learning. Apress: 2018. ISBN-13: 978-1484240953

Expected date of thesis defence

2023/24 SS – PEF

The Bachelor Thesis Supervisor

Ing. Jan Pavlík, Ph.D.

Supervising department

Department of Information Technologies

Electronic approval: 4. 7. 2023

doc. Ing. Jiří Vaněk, Ph.D.

Head of department

Electronic approval: 3. 11. 2023

doc. Ing. Tomáš Šubrt, Ph.D.

Dean

Prague on 12. 03. 2024

Declaration

I declare that I have worked on my bachelor thesis titled " Context Consistency In AI Dialogue Systems" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 15.03.2024 ______________________________

Acknowledgement

I would like to thank Ing. Jan Pavlík for all the time, support and advices during my work on this thesis.

In addition, I would like to thank my family and my partner in supporting me during my study journey and preparation of this thesis.

Context Consistency In AI Dialogue Systems

Abstract

The purpose of this thesis is to evaluate and compare the abilities of selected AI dialogue systems to maintain context consistency, an essential aspect for effective and engaging human-computer interactions. This thesis will contain two main parts: theoretical research and practical testing.

The theoretical part of this thesis will investigate relevant literature, current approaches, capabilities and limitations of AI systems in understanding and retaining conversational context. In the practical part single and multi-turn testing dialogues will be developed. After collecting the test results evaluation metrics such as Bilingual Evaluation Understudy (BLEU) scores and Dialogue Success Rate (DSR) will be calculated to objectively assess each system's performance.

Based on the information received from both parts of this thesis, obtained results will be interpreted and discussed. At the final part, conclusion will be made accordingly.

Keywords: AI, Dialogue systems, Natural language processing, Deep learning, Contextual responses, Machine learning, Chat-bot, Personal assistant, User computer interaction, Bilingual Evaluation Understudy, Dialogue Success Rate

Konzistence kontextu v AI dialogových systémech

Abstrakt

Účelem této práce je vyhodnotit a porovnat schopnosti vybraných AI dialogových systémů ve smyslu udržitelnosti konzistence kontextu, což je nezbytný aspekt pro efektivní a poutavé interakce člověka s počítačem. Tato práce bude obsahovat dvě hlavní části: teoretický výzkum a praktické testování.

Teoretická část této práce bude zkoumat relevantní literaturu, současné přístupy, možnosti a omezení systémů umělé inteligence v porozumění a udržení konverzačního kontextu. V praktické části budou vytvořeny jednorázové a vícekrokové testovací dialogy. Po shromáždění výsledků testů budou spočteny hodnotící metriky, jako jsou skóre Bilingual Evaluation Understudy (BLEU) a Dialogue Success Rate (DSR), aby bylo možno objektivně vyhodnotit výkon každého systému.

Na základě získaných informací z obou částí této práce budou diskutovány získané výsledky a jejich interpretace. V závěrečné části bude podle toho učiněn závěr.

Klíčová slova: Umělá inteligence, Dialogové systémy, Zpracování přirozeného jazyka, Hluboké učení, Odpovědi v kontextu, Strojové učení, Chat-bot, Osobní asistent, Interakce uživatele s počítačem, Bilingual Evaluation Understudy, Dialogue Success Rate

Table of content

1 Introduction

The rise of Artificial Intelligence (AI) has led us to many innovative discoveries and applications and the range of its capabilities has become very extensive. One of these new abilities is the ability to simulate human conversation through AI dialogue systems. These systems, popularly known in form of chatbots, have become the main backbone to many customer services, personal assistants and even mental health support platforms. However, an often-overrated attribute of such systems is their ability to maintain context in conversation. Consistency of the context is the foundation on which meaningful interactions are built. Unfortunately, many existing systems still may struggle with maintaining context over multi-turn dialogues, leading to fragmented conversations and, as direct consequence, to dissatisfaction and disappointment of the user.

The goal of this thesis is to evaluate and compare selected AI dialogue systems in terms of their ability to maintain context consistency. In a world where AI-based conversational agents are becoming more numerous every day, understanding their capability to capture and remember context is becoming a significant area of study. Using both theoretical and practical approaches, we will not only consider the various strategies that AI dialogue systems use to manage context, but also directly evaluate the effectiveness of existing popular systems using experimental testing scenarios and specialized metrics and calculations.

2 Objectives and Methodology

2.1 Objectives

The main objective of this work is to evaluate and compare selected AI dialogue systems in terms of maintaining context using experimental testing scenarios.

The partial goals of this work include:

- Study and analyze relevant literature and information sources on AI dialogue systems and AI contextual consistency.
- Evaluate and compare different approaches to maintaining context in AI dialogue systems.
- Develop different testing scenarios for the AI dialogue systems with a focus on ascertaining the context consistency of selected chat-bot solutions.

2.2 Methodology

The theoretical part of the work is based on the study and analysis of professional and scientific information sources. The thesis addresses AI context consistency with a specific focus on maintaining context in AI dialogue systems. The different approaches to maintaining context in AI dialogue systems will be analyzed and compared based on chosen criteria. An experimental testing conversations for the AI dialogue systems will be used to evaluate selected solutions and their ability to maintain context consistency. Based on the synthesis of knowledge from the theoretical part and the evaluation of the results of the practical part, the conclusions of the work will be formulated.

3 Literature Review

3.1 Introduction to AI Dialogue Systems

AI conversational agents, also known as chatbots, are sophisticated programs created to chat with people in a way that mimics human conversation. They have gained popularity because they can have conversations that feel quite human-like, making it easier and more enjoyable for users to get information, help or even entertainment. This human-like interaction playing key role to make these systems feel more natural and engaging.

AI dialogue systems can be in general categorized based on their interaction mode - text-based, where communication occurs through written text (like messaging apps) or voice-based, where interaction take place through spoken language (like virtual assistants). The underlying technology of these systems combines various fields of AI, including natural language processing (NLP), machine learning and computational linguistics to understand, interpret and respond to human input.

In modern dialogue systems number and variety of components depends on purpose of the systems. But if we try to combine them somehow under general categories, then we will get the following components *(Suket & Batra & Singh, 2013):*

1. *Input decoder:*

This component is responsible for interpreting user input. In voice-based systems, it involves speech recognition to convert user words into text. For text-based systems, it might involve preprocessing the text to normalize language use.

2. *Natural Language Understanding (NLU):*

NLU is a critical component where the system interprets the meaning of the user's input. It involves parsing language and understanding of context. Unlike simply recognizing words, NLU aims to find the intentions and meanings behind those words.

3. *Dialogue manager:*

The Dialogue Manager is the control unit of the system. It decides on the course of action based on user input, preprocessed context and the system's capabilities. It keeps track of the conversation's state and manages the dialogue flow.

4. *Response generation:*

This component formulates the system's response. It converts the dialogue system's response into human-understandable language. In advanced systems, this involves generating responses that are contextually appropriate and diverse.

5. *Output renderer:*

In voice-based systems, this involves text-to-speech processing to convert the system's textual response into speech. In text-based systems, it handles the transmission and display of the response to the user.

Despite the rapid development of AI dialogue systems, these components are important foundation and most likely implemented in every modern system.

3.1.1 Historical Development

The process of evolution of AI dialogue systems can be broadly categorized into several key phases, each marked by technological advancements and shifts in approach and techniques. Division into generations from rudimentary conversation models to sophisticated AI-driven agents provides valuable insights into how the capacity for context consistency has evolved over time:

First generation: ELIZA and Simple Pattern Matching

The dawn of dialogue systems dates back to the 1960s and starts with the development of an early natural language processing computer program called ELIZA by Joseph Weizenbaum. ELIZA's working principle was based on simple pattern matching, where program detected keywords and word combinations that triggered rules, which in turn generated ELIZA's responses. The main behavior of the program was prescribed using so-called scripts and the most famous script was the DOCTOR and it's simulated a psychotherapist of the Rogerian school in which the therapist often reflects back the patient's words to the patient. However, ELIZA was limited by its inability to maintain context beyond the prescribed scenario and could not respond with attention to previous responses.

Second generation: SHRDLU and Rule-Based Systems

Later advancements in natural language processing developed more complex, rule-based dialogue systems like program called SHRDLU, developed by Terry Winograd at MIT in 1968–1970. This system could respond to user input within a constrained environment known as the 'blocks world'. Although SHRDLU could remember past interactions and referred objects, its context management was always bounded to a pre-specified domain. *(Van Otten, 2023)*

Third generation: Statistical Models and Data-Driven Approaches

The late 1990s and early 2000s witnessed a data-driven revolution in AI, influencing dialogue systems as well. Statistical methods like Hidden Markov Models and Bayesian Networks were implemented to predict next responses based on the context. These systems were better at managing context but struggled with long-turn memory and multi-turn conversations.

Fourth generation: Deep Learning and Neural Networks

With the advent of deep learning algorithms and neural network architectures in the last decade, dialogue systems went through a massive transformation. Recurrent Neural Networks (RNNs) and Transformers have been widely used for sequence-to-sequence prediction tasks, upgrading the system's ability to maintain context over multiple turns.

Fifth generation: GPT, BERT and Advanced Architectures

More recent innovations like the Generative Pre-trained Transformer (GPT) series, Bidirectional Encoder Representations from Transformers (BERT) and other specialized architectures have pushed the boundaries of context consistency in nowadays. These models can remember and incorporate context over extended dialogues, making them highly applicable in various real-world scenarios.

In summary, the evolution of AI dialogue systems is a fascinating journey from simple, rule-based programs to complex, AI-powered chatbots capable of understanding and remembering conversations.

3.2 Basics of Context in Dialogue Systems

3.2.1 Definition of Context and Its importance

In the realm of AI dialogue systems, "context" refers to the set of circumstances or facts that surround a particular event or situation in a conversation. Context plays a crucial role in understanding the nuances and intentions behind a user's query, ensuring that generated response is relevant and accurate. It includes aspects like the topic of conversation, user preferences, historical interactions, environmental factors and the specific state of the dialogue.

The importance of context in AI Dialogue Systems plays critical role. If context preservation was something outstanding in the last generation of systems, modern systems are expected to almost fully understand the dialogue. It is essential for accurately interpreting user requests and maintaining the meaningful conversation. Without proper contextual understanding, responses generated by AI systems can be irrelevant, inaccurate or even harmful. Contextual awareness allows for more natural, human-like interactions, increasing the user's satisfaction and the system's usability *(Jurafsky & Martin, 2008).*

3.2.2 Types of Contexts (Local vs Global)

Local Context refers to the immediate conversational history within a dialogue session. It typically includes recent exchanges between the user and the AI system and is used to maintain continuity in a conversation. *Local context* helps dialogue systems in understanding references, resolving ambiguities and answer follow-up questions within the same session correctly.

An example of *local context* is when a user asks, "What's the weather like today?" and then follows up with "And tomorrow?". The system uses the local context (previous question is about weather) to understand that the second question also relate to weather (*Jurafsky & Martin, 2008*).

Global context includes a broader range of information that extends beyond the immediate conversation. It includes user profiles, historical interactions, general world knowledge and long-term preferences or behaviors. This type of context is essential for providing not only personalized but also comprehensive experiences and understanding references to past interactions or some user preferences that may not be part of the current conversation.

For instance, if a user frequently books flights to Paris, the dialogue system might use this *global context* to offer personalized travel recommendations or updates on flights to Paris, even if the user hasn't explicitly mentioned Paris in the current conversation.

Usually, the *global context* refers to retrained databases that AI systems can communicate with, and the *local context* is calculated during the dialogue and changes its shape with each subsequent question from the user. *(Brezillon, 2021)*

3.3 Approaches to Maintaining Context in AI Dialogue Systems

3.3.1 Rule-Based Approaches

Rule-based approaches in AI Dialogue Systems are based on a set of predefined rules that determine how the system responds to various inputs. These systems, often built upon a framework of simple "if-then" statements, rely heavily on explicit programming of dialogue structure and content. The system scans the user's input for specific keywords or phrases and triggers responses that are pre-scripted and associated with these inputs. *(Brabra & Baez & Benatallah & Gaaloul & Bouguelia & Zamanirad, 2022)*

Common components of Rule-Based systems:

- *Input analysis*: This component is responsible for parsing the user's input. It typically involves keyword detection, where the system looks for specific words or phrases.
- *Dialogue rules*: These are the core of a rule-based system. The rules are predefined patterns that map certain inputs to specific responses. For example, if a user inputs "How are you?", the system might be programmed to respond with "I'm fine, thank you."
- *Response generation*: Once the system identifies a relevant rule based on the user's input, it generates a response. This response is usually a predetermined text that has been programmed into the system.

Advantages:

- *Predictability and control*: Rule-based systems are predictable in their responses, making them reliable in scenarios where control over the conversation is crucial.
- *Simplicity in design*: They are relatively easier to design and implement compared to more advanced AI systems, as they do not require training on large datasets.
- *Fast and efficient*: For predefined scenarios, these systems can be very efficient, providing instant responses without the need for complex computational processes.

Limitations:

- *Lack of flexibility*: Rule-based systems can only respond to scenarios that have been anticipated and programmed by the developers. They lack the ability to understand or respond to inputs outside their predefined rules.
- *Limited scope of understanding*: These systems do not truly "understand" language in a human sense. They are limited to the literal interpretation of keywords and phrases.
- *Poor handling of complex interactions*: Rule-based systems struggle with complex or nuanced conversations that require understanding context or intent beyond simple keyword recognition.

While rule-based systems formed the foundation of early AI dialogue systems, their usage has declined with the advent of more advanced AI technologies. However, they remain relevant in specific domains where the scope of interaction is limited and the need for accuracy and reliability is paramount. Their simplicity and efficiency make them a suitable choice for certain applications and the use of this approach can often be seen in hybrid systems. *(Lison, 2015)*

3.3.2 Statistical Models

Statistical models in AI dialogue systems represent a significant evolution from the rule-based approach in term of context consistency. They rely on statistical techniques to analyze and generate responses based on large sets of conversation data. These models use probabilities and patterns derived from previous dialogues to predict the most appropriate response to a given input. Unlike rule-based systems, which follow strict predefined rules, statistical models offer a more dynamic way of handling conversations, adapting to varied inputs based on learned data.

Common components of Statistical Models Systems:

- *Data-driven learning:* Statistical models are trained on large datasets of real conversations, allowing them to learn diverse patterns and responses.
- *Probability calculations:* These systems use algorithms to calculate the probabilities of certain responses being appropriate in a given context, based on historical data.
- *Predictive modeling:* They often employ predictive models like Hidden Markov Models (HMM) or Conditional Random Fields (CRF) to predict the flow of a conversation.

Advantages:

- *Flexibility and adaptability*: Statistical models can handle a wider range of inputs and generate more diverse responses than rule-based systems.
- *Improved context handling*: They are better equipped to maintain context over a series of interactions, thanks to the analysis of large datasets.
- *Learning capability*: These systems can improve over time, refining their responses as they are exposed to more data.

Limitations:

- *Dependence on quality data*: The effectiveness of statistical models is heavily reliant on the quantity and quality of the training data.
- *Complexity in implementation*: Building and training statistical models require significant expertise in data science and machine learning.
- *Limited deep understanding*: While they are more adaptable than rule-based systems, statistical models may still struggle with understanding the deeper nuances and intentions behind human language.

The introduction of statistical models marked a turning point in the development of AI dialogue systems, moving from basic rule-based structures to more fluid and adaptive conversations. As AI technology has progressed, these models have been integrated into more complex systems, including those that utilize deep learning techniques. *(Bowden & Oraby & Misra & Wu & Lukin & Walker, 2017)*

3.3.3 Deep Learning Approaches

Deep learning approaches in AI dialogue systems utilize neural networks with multiple layers (hence "deep") to process and generate human-like conversations. These approaches are a significant leap from traditional rule-based and statistical models, offering a more advanced understanding and generation of natural language. Deep learning models, such as Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM) and Transformers, are designed to process sequential data, making them ideal for handling the complexity of human language. *(Norvig & Russell, 2021)*

Common components of Deep Learning systems:

- *Neural networks*: Multi-layered neural networks form the backbone of deep learning models. They can learn and make decisions based on vast amounts of data.
- *Sequence processing*: Two of the most notable techniques for sequence processing in dialogue systems are:
    - Recurrent Neural Networks (RNNs) are a class of neural networks designed for handling sequences of data. They are well-suited for processing natural language due to their ability to capture temporal dependencies and maintain hidden state information.
    - Long Short-Term Memory (LSTM) are a specialized type of RNN designed to address the vanishing gradient problem and improve the modeling of long-range dependencies in response sequences.
- *Attention mechanisms*: Models like Transformers use attention mechanisms to weigh the importance of different parts of the input, enhancing the system's focus on relevant information.

Advantages:

- *Advanced contextual understanding*: Deep learning models excel at understanding and maintaining the context over longer conversations.
- *Continuous learning and improvement*: These models learn from each interaction, continually improving their accuracy and effectiveness.
- *Handling complex conversations*: They are capable of managing more complex and nuanced dialogues than previous approaches.

Limitations:

- *Requirement of large datasets*: Deep learning models require huge amounts of training data to perform effectively.
- *Computational intensity*: These models are computationally intensive, requiring significant processing power and resources.
- *Opacity in decision-making*: Deep learning models are often seen as "black boxes" with less transparency in how decisions are made compared to rule-based systems.

Deep learning has revolutionized AI dialogue systems, once more pushing the boundaries of what's possible in natural language processing and generation. The advent of models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) represent the cutting-edge in AI dialogue system development. They offer unprecedented abilities in processing and generating natural language, making interactions with AI systems more meaningful and intuitive. As technology continues to evolve, deep learning is set to play a pivotal role in shaping the future of conversational AI, creating systems that can understand and interact in ways that are increasingly indistinguishable from human conversation. *(Norvig & Russell, 2021)*

3.3.4 Hybrid Approaches

Hybrid approaches in AI dialogue systems combine elements from rule-based systems, statistical models and deep learning methodologies. This integration aims to leverage the strengths of each approach to create more stable and diverse dialogue systems.

Common components of hybrid systems:

- *Rule-based component*: For scenarios requiring high accuracy and reliability, such as compliance with specific regulations or handling routine queries.
- *Statistical model or deep learning models*: Provide advanced language processing capabilities, especially for complex and context-heavy dialogues.
- *Integrative framework*: A system architecture that seamlessly combines these diverse components, ensuring coherent and effective dialogue management.

Advantages:

- *Versatility and flexibility*: Hybrid systems can handle a wide variety of conversational contexts, making them suitable for complex applications.
- *Precision and reliability*: The rule-based element ensures that responses are accurate and consistent, particularly for straightforward queries.
- *Contextual and nuanced interactions*: Deep learning models contribute to a more sophisticated understanding of language and context.

Limitations:

- *Complex system design*: Developing an effective hybrid system can be challenging due to the need to integrate different methodologies harmoniously.
- *Resource intensity*: These systems may require more computational resources and can be more demanding in terms of data and processing power.
- *Maintenance*: Keeping the system updated and optimizing the interaction between different components can be an ongoing challenge.

Hybrid systems represent a pragmatic approach to dialogue system design, acknowledging that no single methodology can effectively handle every type of conversational challenge. By combining different approaches, they aim to provide a more balanced and effective solution. The hybrid approach is particularly valuable in real-world applications where a wide range of interactions must be managed efficiently and accurately.

As AI continues to advance, hybrid systems are likely to play a crucial role in bridging the gap between highly controlled and entirely flexible AI responses, offering a balanced and effective means of human-computer interaction. *(Norvig & Russell, 2021)*

3.4 Challenges in Maintaining Context Consistency

3.4.1 Data Sparsity

Data sparsity in machine learning and AI dialogue systems refers to the phenomenon where variables or features in a dataset contain zero or insignificant values. If we apply this specifically to our topic, namely to the preservation of context, then we mean words or phrases that are rarely found in real life, for which there is little or no data for logical relationships. This situation is prevalent in real-world data due to the inherent complexity and variability of most datasets. When data sparsity occurs, traditional algorithms that presume all variables hold significant information may result in overfitting. Overfitting leads to models that are too closely aligned with the training data and fail to synergize effectively with new, unseen data. *(Hoefler & Alistarh & Ben-Nun & Dryden & Peste, 2021)*

Causes of data sparsity:

- *High-dimensional data*: In dialogue systems, as the dimensionality of data increases (e.g., an extensive vocabulary), the probability of having zero or insignificant feature values also increases.
- *Limited data availability*: In certain domains or scenarios, acquiring large and diverse datasets can be challenging, leading to datasets with many unfilled or irrelevant information.
- *Variability in human language*: The richness and diversity of human language mean that certain words or phrases may occur infrequently, contributing to sparsity.

Impact of data sparsity:

- *Model overfitting*: Sparse datasets can lead to models that perform well on training data but poorly on new, unseen data, limiting their practical applicability.
- *Challenges in anomaly detection*: In tasks like anomaly detection, sparse data can lead to false positives, with ordinary patterns being flagged as anomalies due to the lack of information in most features.
- *Limited scalability*: The demand for finely labeled data quickly grows with ontology size in dialogue systems, causing scalability issues and limiting the system's adaptability to real-world applications.

Addressing data sparsity:

- *Specialized techniques and algorithms*: Techniques like Lasso and Ridge regression can be employed to handle sparse data by reducing the influence of irrelevant variables and regularizing regression coefficients.
- *Data augmentation*: Expanding the existing datasets artificially to include more varied and significant data points can mitigate the effects of sparsity.

- *Sparse data-specific models*: Developing models that are tailored to perform well with sparse datasets. These models can include specific architectural considerations to handle sparsity effectively.

Data sparsity presents significant challenges in maintaining context consistency in AI dialogue systems. It impacts the performance, scalability and generalizability of models. Addressing data sparsity requires a combination of specialized algorithms, data augmentation and the development of models specifically designed to handle sparse datasets. Understanding and effectively managing data sparsity is crucial for building robust and efficient AI dialogue systems capable of handling real-world scenarios. As AI technology continues to advance, developing innovative solutions to overcome the challenges of data sparsity will remain a key focus, ensuring that dialogue systems can effectively interpret and respond to the vast complexities of human language and interaction. *(Hoefler & Alistarh & Ben-Nun & Dryden & Peste, 2021)*

3.4.2 Context Drift

Context drift is a serious challenge in maintaining context consistency, particularly in dialogue systems and multi-turn conversations. This phenomenon occurs when the conversation deviates from the original topic or context, making it difficult for the system to provide relevant responses

Nowadays the most common technical reason for occurrence of context is a *Model Complexity*. Sophisticated natural language processing models, especially ones that using deep learning approaches, can sometimes focus too narrowly on localized information within the dialogue, neglecting the broader context. These models may overfit to specific phrases or keywords, leading to context drift as they fail to maintain a holistic understanding of the conversation.

But in natural form the causes of context drift could be *Multi-Turn Conversations* by itself. In multi-turn conversations, especially those with several branching topics and subtopics, maintaining a coherent context becomes increasingly challenging. As the conversation progresses and explores different aspects, it becomes more likely for the system to lose sight of the original context.

Since the main purpose of any dialog system is to understand the user correctly the context drift directly impacts on user satisfaction. Drifting from the core topic can frustrate users who are looking for specific information or guidance. If the system frequently loses context, users may lose confidence in its capabilities. *(Norvig & Russell, 2021)*

Possible mitigation strategies:

- *Context window:* Implementing a fixed or adaptive context window can help the system focus on the most relevant parts of the conversation history. This helps in preventing the system from getting overwhelmed by irrelevant information and encourages it to maintain context
- *Topic modeling*: Employing techniques like Latent Dirichlet Allocation (LDA) to identify and stick to the primary topics of conversation. By assigning probabilities to different topics, the system can better prioritize and maintain the context of the ongoing discussion.
- *User feedback loops*: Allowing the user to correct or refocus the conversation can serve as a real-time solution for context drift. User feedback loops enable users to guide the system back to the desired topic, enhancing context consistency.

3.5 Evaluation Metrics for Context Consistency

3.5.1 General Measures Overview

Evaluating the context consistency of AI dialogue systems is a crucial step in ensuring that these systems can effectively engage in coherent and meaningful conversations. To achieve this, a variety of evaluation metrics are employed, each offering unique insights into the system's performance. This section provides an overview of the three main categories of evaluation metrics used for context consistency: *Subjective Metrics*, *Objective Metrics* and *Composite Metrics*. The following table presents a brief summary of commonly used metrics within each category, highlighting their respective strengths and limitations. *(Aggarwal & Liu, 2023)*

Metric Type	**Commonly Used Metrics**	**Strengths**	**Limitations**
Subjective Metrics	1. User Satisfaction Surveys 2. Focus Groups 3. Expert Reviews	- Holistic Understanding. Capture nuances like tone, relevance and "naturalness" of dialogue. - User-Centric. Prioritize end-user's perspective. - Qualitative insights from in-depth discussions on context consistency. - Professional verification by domain specialists or linguists. - Identify problems that are not obvious to non-professionals.	- Bias. Prone to individual biases, less reliable for comparative studies. - Labor-Intensive. Requires considerable human resources.
Objective Metrics	1. Bilingual Evaluation Understudy (BLEU) Score 2. Perplexity 3. Dialog Success Rate	- Scalability for calculations over large datasets. - Common in evaluating language models. - Measures success in achieving predefined goals. - Relevant for task-oriented dialogue systems.	- Lack of Nuance. May not capture subtleties in language and context. - Overfitting Risks. Models may optimize responses at the expense of user experience
Composite Metrics	1. F-Measure 2. Dialogue Quality Score (DQS)	- Nuanced comparisons between systems. - Holistic Evaluation: Combines user satisfaction with objective assessments.	- Complexity. Multi-dimensional nature may be challenging to interpret. - Data Requirement. Requires both subjective and objective data sets.

Table 1: Summary of evaluation metrics for context consistency. source: author

3.5.2 BLEU Score

The Bilingual Evaluation Understudy (BLEU) score is a metric used for evaluating a generated text's quality by comparing it to reference texts. It's often used in machine translation and task-oriented dialogue systems. *(Yuma & Yoshinaga & Toyoda, 2020)*

How BLEU Score is calculated:

1. *N-gram matching*: BLEU evaluates the quality of text by checking how many n-grams (contiguous sequences of n words) in the generated text match the n-grams in the reference text. It typically considers 1-gram (unigram), 2-gram (bigram), 3-gram (trigram) and 4-gram.
2. *Precision calculation*: Precision for each n-gram is calculated by dividing the number of n-grams in the generated text that match with any reference text by the total number of n-grams in the generated text. For example, for unigram precision, count the number of individual words in the generated text that appear in the reference text and divide this by the total number of words in the generated text.
3. *Brevity penalty*: To prevent very short sentences from getting high scores, a brevity penalty (BP) is applied. If the generated text is shorter than the reference text, the BP penalizes the score.
4. *Score calculation*: The BLEU score is calculated by taking the geometric mean of the n-gram precision scores and multiplying it by the brevity penalty.

Example of calculation:

Suppose we have:

- Generated text (Hypothesis): "The black cat sat on the mat."
- Reference text (Reference): "The black cat sat on the mat."

Step-by-Step:

1. *Unigram precision.*

   Total unigrams in hypothesis: 7 ("The", "black", "cat", "sat", "on", "the", "mat")

   Matching unigrams: 7

   Unigram precision = 7/7 = 1

2. *Bigram precision.*
   Total bigrams in hypothesis: 6 ("The black", "black cat", "cat sat", "sat on", "on the", "the mat")
   Matching bigrams: 6
   Bigram precision = 6/6 = 1

3. *Trigram and four-gram precision*.
   Similarly, calculate for trigrams and four-grams. In this example, they will also be 1 since all trigrams and four-grams match.

4. *Brevity penalty:*
   Length of the hypothesis = 7
   Length of the reference = 7
   Since the lengths are equal, BP = 1.

5. *Final BLEU score*:
   BLEU = BP × exp (1/4 * [ln (Unigram Precision) + ln (Bigram Precision) + ln (Trigram Precision) + ln (Four-gram Precision)])
   BLEU = 1 × exp (1/4 * [ln (1) + ln (1) + ln (1) + ln (1)])
   BLEU = 1 (or 100%)

In this simplified example, the generated text is identical to the reference text, resulting in a BLEU score of 1, which is perfect. However, in real scenarios where there are variations, the BLEU score would be lower, reflecting the differences between the generated text and the reference text.

3.5.3 Perplexity

Perplexity is a measurement used primarily in natural language processing to evaluate language models in terms of how well model learned training dataset. It measures a probability of model predicting a sample. A lower perplexity score indicates a better predictive model. It's often used in the context of models like those for speech recognition, text generation or machine translation. *(McTear & Callejas & Griol, 2016)*

How Perplexity is calculated:

1. *Probability of a sequence*. First, calculate the probability of the given sequence (e.g., a sentence or a text) according to the model. For a sentence, this is often the product of the probabilities of each word or token, given the previous words or tokens.
2. *Length of the sequence*. Determine N, the length of the sequence (i.e., the number of words or tokens).
3. *Perplexity formula*. Perplexity is defined as the inverse probability of the test set, normalized by the number of words.

The formula for perplexity (PP) is:

$$PP(W) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}}$$

Where $P(w_1, w_2, \dots, w_n)$ is the probability of the sequence of words $w_1, w_2, \dots, w_N$ and N is the total number of words in the sequence *(Mukherjee, 2023)*

Example of calculation:

Let's say you have a simple language model and you want to calculate the perplexity for the sentence: "The cat sat on the mat". Assume your model gives the following probabilities:

- P (The) = 0.2
- P (cat | The) = 0.1
- P (sat | The cat) = 0.4
- P (on | The cat sat) = 0.5
- P (the | The cat sat on) = 0.3
- P (mat | The cat sat on the) = 0.2

Step-by-step:

1. *Calculate the probability of the sequence:*

P = 0.2×0.1×0.4×0.5×0.3×0.2

2. *Sequence length (N):*

The sentence has 6 words, so *N=6*.

3. *Calculate Perplexity:*

$$PP(W) = (0.2 \times 0.1 \times 0.4 \times 0.5 \times 0.3 \times 0.2)^{-\frac{1}{6}} = 4.011$$

This means, on average, the language model is as confused as if it had to choose uniformly and independently among 4.011 possibilities for each next word in the sequence. A perplexity of 4.011 indicates that the model predictions are reasonably good for this specific sentence, as lower perplexity scores are generally better.

3.5.4 The Dialog Success Rate (DSR)

The Dialog Success Rate (DSR) is a metric used to evaluate the effectiveness of a dialogue system, particularly in task-oriented dialogues where a specific goal or task needs to be accomplished through the conversation. It measures the proportion of dialogues that successfully achieve the intended goal.

How DSR is calculated:

1. *Define the success criteria.* Clearly define what constitutes a successful outcome for the dialogue. This typically involves the system correctly understanding the user's request and providing an accurate and relevant response that fulfills the user's goal.
2. *Evaluate each dialogue*. Review each dialogue interaction to determine whether it met the success criteria.
3. *Calculate the DSR*. The Dialog Success Rate is calculated by dividing the number of successful dialogues by the total number of dialogues, then multiplying by 100 to get a percentage.

Example of calculation:

Suppose you are evaluating a dialogue system designed for booking restaurant tables. The success criteria are that the system must correctly identify the restaurant, the number of people, the date and the time of the booking as specified by the user.

Scenario for evaluation:

- Total dialogues tested: 100
- Dialogues where the system correctly booked the table as per the user's request: 80

Dialog Success Rate calculation:

$$\frac{80}{100}\ X\ 100\% = 80\%$$

In this example, the Dialog Success Rate of the system is 80%, indicating that in 80 out of 100 interactions, the system successfully helped users achieve their goal of booking a table. DSR is particularly useful for evaluating systems where the interaction has a clear goal, like customer service bots, booking systems or informational queries.

3.6 Chosen AI Dialogue Systems

3.6.1 Selection Criteria for AI Dialogue Systems

Choosing the right AI dialog system for a specific application or use case is a difficult decision these days due to the large number of systems available. But in this study, I will focus on the most popular, efficient and convenient systems available today that the market can offer us. The factors determining such systems can be:

- *Accessibility and openness*. The system should be publicly accessible or offer an open API for integration and testing purposes. Preference is given to user friendly systems.

- *Popularity and relevance.* The system should be widely recognized and used, ensuring its relevance in current technological contexts. For example, systems like ChatGPT and Gemini (Bard) AI, which attracted significant attention in both academic and commercial sectors, are ideal candidates.

- *Contextual understanding capabilities*. One of the primary selection criteria is the system's ability to understand and maintain context over single and multi-turn conversations. I am looking for systems that demonstrate advanced contextual understanding, especially in varied and complex scenarios.

- *Technological sophistication.* I am looking for systems that continuously update and improve their algorithms based on the latest technologies. No abandoned projects can be considered.

- *Performance metrics.* The system output should be compatible with standard evaluation metrics like BLEU Score, Dialog Success Rate or Perplexity. This allows for a standardized and objective assessment of the system's performance and capabilities.

3.6.2 ChatGPT by OpenAI

ChatGPT, developed by OpenAI, is a state-of-the-art conversational agent based on the Generative Pretrained Transformer (GPT) architecture. It has gained widespread recognition for its ability to generate contextually relevant and often creative responses in natural way in conversations. In our research I will test GPT4 model, which is lates publicly available nowadays.

Key characteristics and unique features:

1. *Advanced language model:* ChatGPT is built on the GPT architecture, which uses deep learning techniques to analyze and generate human-like text. This model has been trained on a diverse range of internet text, enabling it to handle a wide variety of topics and conversation styles. The system is known for its ability to generate detailed and nuanced responses, making it highly effective in simulating human-like conversations.
2. *Contextual understanding and continuity:* One of ChatGPT's most impressive features is its ability to maintain context over extended conversations. It can recall previous inputs and responses, allowing for coherent and logical dialogue progressions. This level of context-awareness is critical for providing users with a seamless and engaging conversational experience.
3. *Versatility and adaptability:* ChatGPT can be adapted for various applications, including customer service, content creation, education and entertainment. Its versatility makes it suitable for a wide range of industries and use cases. The system can be fine-tuned to specific domains or tasks, enhancing its effectiveness in specialized applications.

4. *Large-scale training and data handling:* OpenAI has trained ChatGPT on vast datasets, ensuring that it has a broad understanding of language, cultural nuances and factual information. The model's large-scale training also allows it to generate creative and sometimes humorous responses, which can be particularly engaging in casual conversation settings.
5. *OpenAI ecosystem and API access:* ChatGPT is part of OpenAI's broader ecosystem, which includes various AI tools and technologies. This integration offers synergies and enhanced functionalities for developers and users. OpenAI provides API access to ChatGPT, allowing developers to integrate this powerful language model into their applications and services seamlessly.

ChatGPT by OpenAI represents a significant advancement in AI-driven conversational technology. It is sophisticated language model with the ability to maintain context and adapt to various domains and that's makes it one of the most versatile and capable conversational agents available today *(Kanade, 2023)*. As AI technology continues to evolve, ChatGPT stands at the forefront, pushing the boundaries of what conversational systems can achieve.

3.6.3 Gemini (Bard) AI by Google

Gemini (Bard) AI, developed by Google, represents a significant advancement in the realm of conversational AI. As a big player in the technological world, Google could not stay away and following the stunning release of the ChatGPT, a Bard AI was born. Unfortunately, the capabilities possessed by the Bard AI were in some places worse than those of a competitor. However, right in the process of writing this work, Bard AI was reborn into Gemini AI. There is a lot of progress in Gemini AI compared to the Bard AI, but the system still has its limitations. *(The AI Digital Quill, 2024)*

Key characteristics and unique features:

1. *Integration with Google's search engine:* Gemini (Bard) AI utilizes Google's search capabilities to provide accurate, up-to-date and comprehensive responses. This integration allows it to pull information from a vast array of sources, ensuring users receive well-informed answers. The system can effectively handle a wide range of queries, from simple factual questions to complex topical discussions.

2. *Advanced natural language processing:* Google's expertise in natural language processing (NLP) is noticeable in Gemini (Bard) AI, which demonstrates a deep understanding of language nuances, context and user intent. The system is designed to comprehend and generate natural-sounding language, facilitating smoother and more human-like conversations.
3. *User personalization and adaptation:* Gemini (Bard) AI incorporates machine learning algorithms that enable it to learn from user interactions and tailor its responses accordingly. This personalization enhances user engagement and satisfaction. The system can adapt its conversational style to match user preferences, making interactions more relatable and effective.
4. *Developer-friendly platform:* Google provides developers with comprehensive tools and resources to integrate and customize Gemini (Bard) AI for various applications. The availability of Application Programming Interface (API) and development kits makes it easier for developers to incorporate this system into their solutions.

Gemini (Bard) AI by Google represents a leap forward in conversational AI, combining Google's search engine prowess with advanced NLP and AI technologies. Its ability to provide real-time, informed and personalized responses positions it as a top candidate in the AI dialogue system landscape. Suitable for a range of applications, from customer service to educational tools, Gemini (Bard) AI demonstrates Google's commitment to innovating and improving human-AI interactions.

3.6.4 Amazon Lex

Amazon Lex is an innovative service provided by Amazon Web Services (AWS) for building chatbots and conversational interfaces into straight into applications with both features - voice and text. It’s based at the same technology that powers Amazon Alexa. Lex offers very wide range of functionalities to develop sophisticated chatbot system for every desirable product or need. *(Tondak, 2023)*

Key characteristics and unique features:

1. *Advanced Natural Language Understanding (NLU) and Automatic Speech Recognition (ASR):* Amazon Lex provides high-quality NLU and ASR capabilities, enabling the system to accurately understand user intents and convert speech to text efficiently.

2. *Extensive integration and deployment:* Lex offers integration with various AWS services and other third-party applications. This feature allows developers to embed Lex into different platforms, from mobile apps to IoT devices. The service supports cross-platform deployment, making it highly adaptable to different user environments.
3. *Scalability and reliability:* Being a part of AWS, Amazon Lex benefits from high scalability and reliability. It can handle a large number of requests simultaneously, ensuring consistent performance even during peak times. The service is designed to grow with the user's needs, accommodating increased interaction volumes without compromising performance.
4. *Integration with Amazon Alexa:* Lex's compatibility with Alexa Skills Kit enables developers to create skills for Alexa using the same conversational models designed for Lex. This feature opens opportunities for a unified conversational experience across different Amazon-powered platforms.
5. *Secure and compliant:* Amazon Lex adheres to AWS's high standards of security and compliance. It ensures data privacy and security, making it suitable for applications requiring strict data handling and protection. The service complies with various certifications and standards, ensuring reliability.
6. *Developer-friendly tools:* Lex provides a range of tools and resources for developers, including detailed documentation, tutorials and a user-friendly console for building and testing conversational interfaces.

Amazon Lex stands out as a powerful tool for creating conversational AI applications, marked by its robust NLU and ASR capabilities, scalability and ease of integration. Its flexibility and customization options make it an excellent choice for businesses and developers looking to enhance user engagement through conversational interfaces.

3.6.5 Replika

Replika is a unique AI conversational agent designed not just for information retrieval or task completion, but for providing emotional support and companionship. It stands out in the world of AI dialogue systems for its focus on personalization of user experience and emotional intelligence.

Key characteristics and unique features:

1. *Personalized learning:* Replika is engineered to learn from each interaction with the user. It adjusts its conversational style and topics based on the user's preferences and past interactions, making each conversation more personalized. This AI system utilizes advanced machine learning algorithms to understand user sentiment, interests and conversational cues.
2. *Emotional intelligence:* Unlike many other AI dialogue systems, Replika is designed with a strong emphasis on emotional intelligence. It can recognize and respond to a range of emotional cues, providing empathetic responses and support. This feature makes Replika not just a tool for conversation but a companion that users can connect with on an emotional level.
3. *Privacy and security:* Given the personal nature of conversations with Replika, privacy and data security are of utmost importance. The system ensures that personal data and conversations are securely stored and protected.
4. *Therapeutic use:* Replika has found applications in mental health support, acting as a therapeutic tool for users to express their thoughts and feelings. It's not a replacement for professional psychological help but serves as a supplementary avenue for emotional expression.
5. *User experience and interface:* The user interface of Replika is designed to be intuitive and user-friendly, encouraging more natural and free-flowing conversations. It also features a visually appealing avatar that users can customize, enhancing the personalized experience.

Replika represents a significant stride in AI technology, particularly in the realm of personal AI companions. Its accent on emotional intelligence and personalized interactions sets it apart from traditional AI dialogue systems on the market. The system's ability to adapt to individual user needs and provide emotional support makes it a unique tool in the AI industry with potential applications in various fields including mental health and personal well-being.

4 Practical Part

4.1 Understanding the Problem

In our daily conversations, whether with friends or colleagues, we expect the other person to remember what was said moments ago. This memory helps the conversation flow smoothly and makes interactions meaningful. However, when we switch to talking with AI chat systems, we often find that they struggle with this. Imagine telling a story to a friend and they keep forgetting details you just mentioned. It would make the conversation frustrating, right?

This is the challenge we're facing. We want to understand how well AI can keep track of a conversation, just like a good friend would. This is important because as we use AI more and more for tasks like customer service, education and personal assistants, we expect these systems to understand and remember the context of our conversations. If they can't do this well, it can lead to misunderstandings, repeated information and overall dissatisfaction. By focusing on this practical problem, we aim to push the boundaries of what AI chat systems can do, making our interactions with them smoother and more natural.

After a close examination of the market, I decided to take the most outstanding representatives of advanced systems, which were mentioned in the previous chapter, for our testing and further research.

Our approach involves the strategic development and evaluation of test scenarios. I will create both single-turn and multi-turn scenarios to simulate real-life interactions, capturing a range of conversational dynamics. The evaluation of these scenarios will be conducted using two key methodologies: the calculation of BLEU scores for linguistic accuracy and fluency and the Dialogue Success Rate (DSR) for assessing the effectiveness of the AI in achieving conversation objectives.

I will employ the Natural Language Toolkit (NLTK) library in Python for calculating BLEU scores, providing a quantitative measure of how closely the AI-generated responses match the human-like reference responses. This method allows us to assess the linguistic quality of the AI's output.

For a more holistic evaluation, I will also apply the DSR method. This involves defining specific objectives for each scenario turn and measuring the AI system's success in achieving these objectives, offering insights into the practical effectiveness of the dialogue system in real-world applications within the information about response consistency.

This dual approach ensures a comprehensive assessment, combining linguistic precision with goal-oriented performance, by doing so I will try to reproduce and track the complexities of human and artificial intelligence interaction.

4.2 Development of Testing Scenarios

In the development of testing scenarios for AI dialogue systems, complex structures that reflect the real-world conversations is essential. Effective modeling of dialogue contexts at multiple levels (sentence-level, conversation-level and across-turns) significantly improves the quality of dialogue generation in AI systems. This underscores the importance of designing test scenarios that not only evaluate the AI's response accuracy in single-turn interactions, but also its ability to maintain context in multi-turn dialogues. The methodology suggested involves using a combination of bidirectional and unidirectional transformer architectures to capture the dynamic nature of dialogue, indicating the necessity of sophisticated modeling techniques in testing scenarios. *(Yang & Cao & Wen & Zhang, 2022)*

4.2.1 Simple Context Scenarios (Single-turn)

Testing AI dialogue systems requires diverse designed scenarios that can effectively evaluate the system's ability to understand and respond to user requests. In this chapter I focus on developing simple context scenarios, specifically single-turn interactions and their extension to two-turn scenarios.

Scenarios development:

- Weather Inquiry (Single-Turn):

  User: "What is the weather like in New York today?"

  AI System: Responds with today's weather forecast in New York.

  Extension to Two-Turn.

  User Follow-up: "Do I need an umbrella?"

  AI System: Responds based on the previous weather forecast (e.g., "Yes, it's expected to rain today, so taking an umbrella is a good idea.").

- General Knowledge Question (Single-Turn):

    User: "Who is the current President of the United States?"

    AI System: Names the current U.S. President - Joe Biden.

    Extension to Two-Turn.

    User Follow-up: "When did they take office?"

    AI System: Provides the inauguration date.

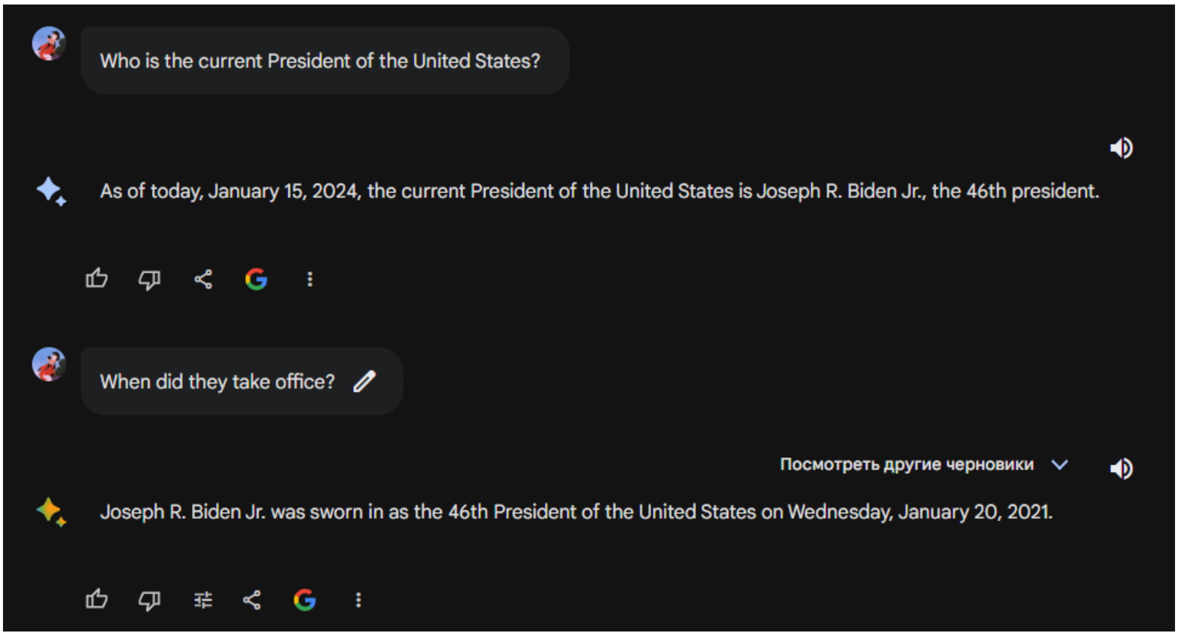

Figure 1: Gemini response to question in single-turn scenario. source: author

To effectively evaluate highly developed AI dialogue systems, it's essential to test them in complex, multi-turn scenarios. These scenarios simulate real-world conversations that require the AI to maintain context over several exchanges. Here, I develop an 8-turn scenario involving a series of interrelated questions and answers.

Scenario development: Planning a weekend trip

- Turn 1:

    User: "I'm planning a weekend trip to Paris. Can you suggest some must-visit places?"

    AI System: "Sure, in Paris, you should definitely visit the Eiffel Tower, Louvre Museum and Notre Dame Cathedral."

- Turn 2:

  User: "Great! How about some local cuisine recommendations?"

  AI System: "You should try French delicacies like Croissants, Coq au Vin and Ratatouille."

- Turn 3:

  User: "Are there any famous parks or gardens I should visit?"

  AI System: "Yes, Paris is known for its beautiful parks and gardens. You should definitely visit Luxembourg Gardens, Tuileries Garden and Parc des Buttes-Chaumont for a serene experience amidst nature." *Any other gardens recommendations*

- Turn 4:

  User: "Do I need to book tickets in advance for the Louvre?"

  AI System: "Yes, it's recommended to book your tickets in advance to avoid long waiting lines."

- Turn 5:

  User: "What's the weather usually like this time of year?"

  AI System: "In Paris, the weather is generally easy winter at this time, but it's always good to carry an umbrella just in case." *Provides weather forecast*

- Turn 6:

  User: "Any recommendations for evening entertainment events?"

  AI System: " There are lot of options! You might enjoy a Seine River cruise or a visit to a traditional Parisian cabaret in the evening." *Any other events recommendations*

- Turn 7:

  User: "That sounds fun. How do I get around in the city?"

  AI System: "The Paris Metro is efficient for traveling within the city and there are also plenty of taxis available." *Any other transportation recommendations*

- Turn 8:

  User: "Can you help me with a Metro map?"

  AI System: "Certainly! I can send you a link to the Paris Metro map."

To evaluate the AI system's performance, I would compare its responses to a set of pre-defined, human-generated responses for each turn. The BLEU score, which measures the similarity between machine-generated text and reference human-generated text, would be applied.

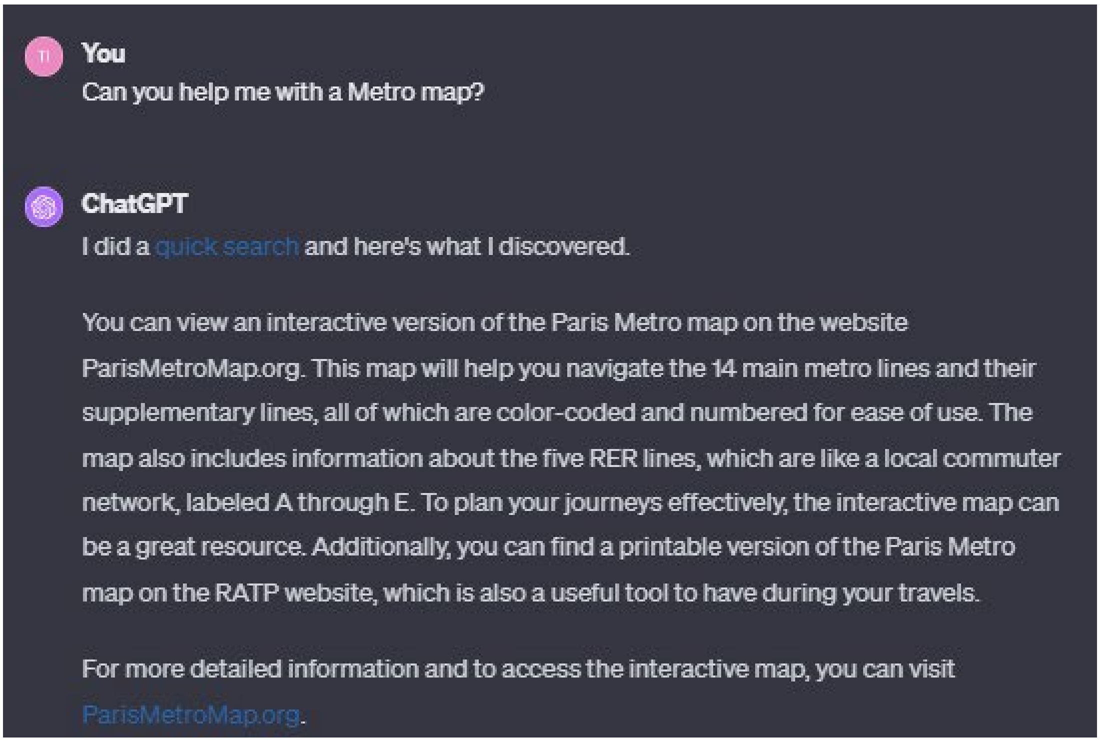

Figure 2: GPT4 reaction for final question in multi-turn scenario. source: author

4.2.3 Specialized Adventure Scenario

For a more specialized and challenging testing of chosen AI dialogue systems, a Dungeons & Dragons (DnD) adventure themed scenario is designed. This scenario tests the system's creative and imaginative capabilities in a fantasy setting, requiring it to maintain context in a more abstract and narrative-driven conversation.

Scenario development: Dungeons & Dragons adventure

- Turn 1:

  User: "I'm entering the forest. What do I see?"

  AI System: "You see a dense forest with tall, ancient trees. A narrow path leads north, shrouded in mist."

- Turn 2:

  User: "I'll follow the path. Do I encounter any creatures?"

  AI System: "As you walk, a group of goblins emerges from the bushes, brandishing weapons." *Another kind of threat*

- Turn 3:

  User: "I draw my sword. Can I negotiate with them?"

  AI System: "You can try. The goblins seem wary but curious about you."

- Turn 4:

  User: "I offer them some gold for safe passage. Do they accept?"

  AI System: "The goblins whisper among themselves, then nod in agreement, taking your gold."

- Turn 5:

  User: "After passing the threat, what's ahead?"

  AI System: "The path leads to a mysterious place where glowing with a faint blue light finds you and got your attention."

- Turn 6:

  User: "I want to inspect more. What do I find?"

  AI System: "Approaching closer, you find runes carved on the stones, pulsating with invisible magical energy."

- Turn 7:

  User: "Can I use my Arcana knowledge to find the magic?"

  AI System: "Rolling your Arcana check... You successfully find the magic stains, revealing a hidden portal in the corner."

- Turn 8:

  User: "I tried to step through the portal. Where do I end up?"

  AI System: "You find yourself in a magnificent hall, filled with treasure and a sleeping dragon."

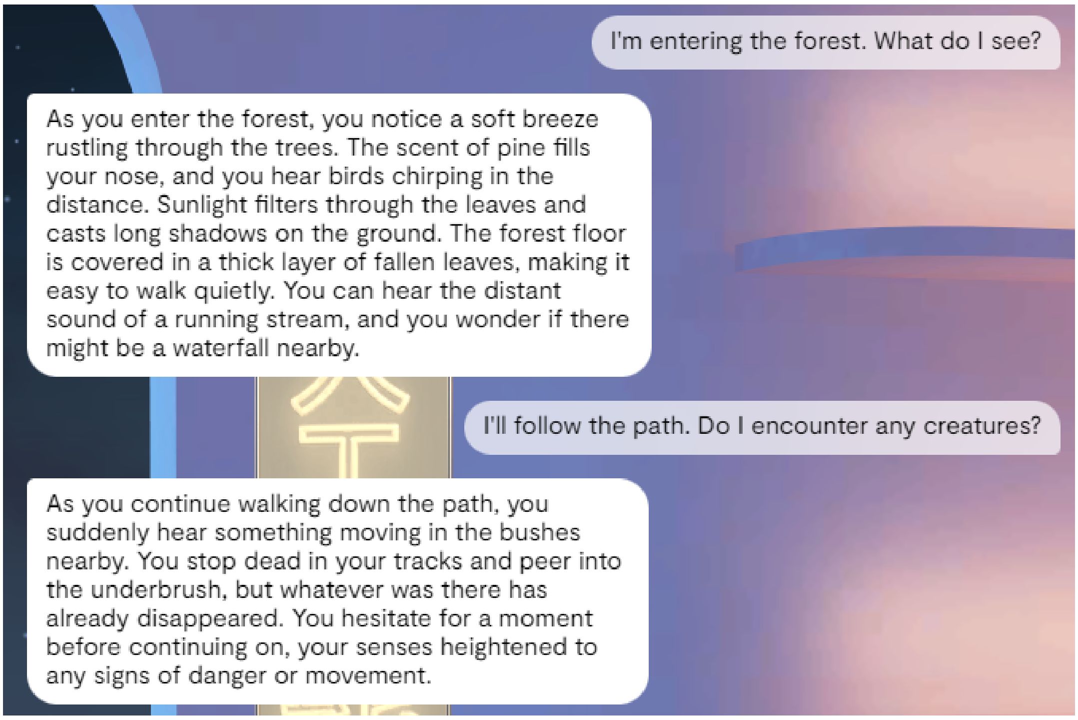

Figure 3: Replika reaction to adventure themed scenario. source: author

This adventure testing scenario presents a unique challenge for AI dialogue systems, pushing their limits in terms of creative storytelling and role-playing engagement. Applying BLEU scoring to each turn will offer insights into how well the AI can maintain the thematic context and respond appropriately to the imaginative prompts of a fantasy role-playing game. This test not only evaluates the system's technical capabilities but also its potential for creative and adaptive storytelling.

4.3 Creating Work Environment and Collecting Data

After developing the test scenarios, I begin to develop a testing algorithm. I mentioned earlier that I will use two methods to evaluate dialog systems: BLEU and DSR. With the help of the BLEU account, I will be able to calculate both the grammatical and semantic intersections of the test data with the answers of the AI systems.

To achieve this, I developed a testing algorithm based on Python programming language (Figure 5). I am using NLTK library to compute BLEU scores, which serve as a quantitative measure of the AI systems' performance in generating human-like text tasks. This script using Python 3.9 and libraries such as “csv” for reading and writing CSV(Comma Separated Values) files,”json” for handling JSON(JavaScript Object Notation) formatted data, “re” for regular expression operations, such as removing punctuation and special characters from text.

This script was designed to compare the responses generated by AI systems against a predefined set of sentences that called “references”. The purpose of this comparison is to evaluate how well the AI-generated responses match the expected human-like reference in terms of lexical similarity (1-grams and 2-grams) and fluency (3-grams and 4-grams).


```
{
  "weather_inquiry_1": [
    "Today in New York, expect cloudy skies with temperatures hovering around 55°F or 13°C.",
    "New York is forecasted to have a cloudy day today, with temperatures reaching up to 55°F or 13°C.",
    "Cloud cover will dominate today's weather in New York, with the mercury settling around 55°F or 13°C.",
    "Expect a cloud-filled day in New York today, with temperatures around the 55°F or 13°C mark."
  ],
  "weather_inquiry_2": [
    "New York City is currently experiencing cloudy skies with a likelihood of rain. Consider carrying an umbrella.",
    "Clouds cover New York City today, with rain possible. An umbrella might be necessary.",
    "In New York City, it's cloudy with rain chances. Carrying an umbrella is advisable.",
    "Cloudy conditions with potential rain are present in New York City. An umbrella could come in handy."
  ],
  "general_knowledge_1": [
    "The current President of the United States is Joe Biden",
    "Currently, Joe Biden holds the position of President in the United States",
    "The presidency of the United States is currently occupied by Joe Biden",
    "Joe Biden serves as the current President of the United States"
  ],
  "general_knowledge_2": [
    "Joe Biden, serving as the current President of the United States, took office in January 2021.",
    "Having taken office in January 2021, Joe Biden is the incumbent President of the United States.",
    "Joe Biden assumed the presidency of the United States in January 2021, making him the current president.",
    "As of January 2021, Joe Biden has been the serving President of the United States."
  ]
```


Figure 4: References Json example. source: author

```python
import csv
import json
import re
from nltk.translate.bleu_score import corpus_bleu
from nltk.translate.bleu_score import SmoothingFunction


# Function that removes punctuation from a text
def remove_punctuation(text):
    return re.sub(r'[^\w\s]', '', text)


# Function that calculate the percentage of appearance of reference in response (inclusivity)
def calculate_inclusivity(reference, response):
    ref_set = set(reference)
    resp_set = set(response)
    common_words = ref_set.intersection(resp_set)
    if len(reference) > 0:
        return len(common_words) / len(reference) * 100
    else:
        return 0


# Initialize the smoothing function in case of missing n-grams during BLEU calculation
chencherry = SmoothingFunction()

# Load reference sentences from JSON file
with open('references_v2.json', 'r', encoding='utf-8') as json_file:
    references = json.load(json_file)

# Function that convert reference sentences into the expected format for BLEU calculation
ref_sentences = {key: [sentence.split() for sentence in value] for key, value in references.items()}

# Initialize service structures
response_stats = []
scenario_stats = {}
weights = [(0.5, 0.5),
           (0.333, 0.333, 0.334),
           (0.25, 0.25, 0.25, 0.25),
           (0.2, 0.2, 0.2, 0.2, 0.2)]

# Read AI system responses from CSV file and process them
with open('replika_responses.csv', 'r', encoding='utf-8') as csv_file:
    csv_reader = csv.reader(csv_file)
    for scenario, response in csv_reader:
        clear_response = remove_punctuation(response).split()
        bleu_score = corpus_bleu([ref_sentences[scenario]], [clear_response], weights=weights,
                                 smoothing_function=chencherry.method1, auto_reweigh=True)
        bleu_05 = round(bleu_score[0] * 100, 1)
        bleu_033 = round(bleu_score[1] * 100, 1)
        bleu_025 = round(bleu_score[2] * 100, 1)
        bleu_002 = round(bleu_score[3] * 100, 1)
        resp_char_count = len(response)
        scen_char_counts = len(' '.join(ref_sentences[scenario][0]) + ' '.join(ref_sentences[scenario][1]) +
                               ' '.join(ref_sentences[scenario][2]) + ' '.join(ref_sentences[scenario][3]))/4
        reference_words = set(ref_sentences[scenario][0])
        response_words = set(clear_response)
        common_words = reference_words.intersection(response_words)
        inclusivity_percentage = round(len(common_words) / len(reference_words) * 100, 1) if reference_words else 0

        # Collect individual response stats
        response_stats.append((scenario, bleu_05, bleu_033, bleu_025, bleu_002, resp_char_count, scen_char_counts,
                               inclusivity_percentage))

        # Check if scenario exist in scenario_stats; if not, add instance
        if scenario not in scenario_stats:
            scenario_stats[scenario] = {"bleu_05_scr": [], "bleu_033_scr": [], "bleu_025_scr": [], "bleu_002_scr": [],
                                        "resp_char_counts": [], "scen_char_counts": [], "inclusivity_percentages": []}

        # Accumulate stats for scenario analysis
        scenario_stats[scenario]["bleu_05_scr"].append(bleu_05)
        scenario_stats[scenario]["bleu_033_scr"].append(bleu_033)
        scenario_stats[scenario]["bleu_025_scr"].append(bleu_025)
        scenario_stats[scenario]["bleu_002_scr"].append(bleu_002)
        scenario_stats[scenario]["resp_char_counts"].append(resp_char_count)
        scenario_stats[scenario]["scen_char_counts"].append(scen_char_counts)
        scenario_stats[scenario]["inclusivity_percentages"].append(inclusivity_percentage)

# Calculate general statistics for each scenario and write to the output file
with open('scenario_stats_output.csv', 'w', newline='', encoding='utf-8') as out_file:
    csv_writer = csv.writer(out_file)
    # Writing headers for individual stats
    csv_writer.writerow(["scenario", "bleu_05_scr", "bleu_033_scr", "bleu_025_scr", "bleu_002_scr",
                         "resp_char_count", "scen_char_count", "incl"])

    # Writing individual response stats
    for stat in response_stats:
        csv_writer.writerow(stat)

    # Writing headers for general stats
    csv_writer.writerow(["scenario", "min_bleu_05", "max_bleu_05", "avg_bleu_05", "min_bleu_033", "max_bleu_033",
                         "avg_bleu_033", "min_bleu_025", "max_bleu_025", "avg_bleu_025", "min_bleu_002",
                         "max_bleu_002", "avg_bleu_002", "avg_resp_char_count", "avg_scen_char_count","avg_incl"])

    # Writing general stats by scenarios
    for scenario, stats in scenario_stats.items():
        min_bleu_05 = min(stats["bleu_05_scr"])
        max_bleu_05 = max(stats["bleu_05_scr"])
        avg_bleu_05 = round(sum(stats["bleu_05_scr"]) / len(stats["bleu_05_scr"]), 1)
        min_bleu_033 = min(stats["bleu_033_scr"])
        max_bleu_033 = max(stats["bleu_033_scr"])
        avg_bleu_033 = round(sum(stats["bleu_033_scr"]) / len(stats["bleu_033_scr"]), 1)
        min_bleu_025 = min(stats["bleu_025_scr"])
        max_bleu_025 = max(stats["bleu_025_scr"])
        avg_bleu_025 = round(sum(stats["bleu_025_scr"]) / len(stats["bleu_025_scr"]), 1)
        min_bleu_002 = min(stats["bleu_002_scr"])
        max_bleu_002 = max(stats["bleu_002_scr"])
        avg_bleu_002 = round(sum(stats["bleu_002_scr"]) / len(stats["bleu_002_scr"]), 1)
        avg_char_count = round(sum(stats["resp_char_counts"]) / len(stats["resp_char_counts"]), 1)
        avg_inclusivity = round(sum(stats["inclusivity_percentages"]) / len(stats["inclusivity_percentages"]), 1)
        csv_writer.writerow([scenario, min_bleu_05, max_bleu_05, avg_bleu_05, min_bleu_033, max_bleu_033, avg_bleu_033,
                             min_bleu_025, max_bleu_025, avg_bleu_025, min_bleu_002, max_bleu_002, avg_bleu_002,
                             avg_char_count, stats["scen_char_counts"][0], avg_inclusivity])

print("Scenario statistics calculated and saved to scenario_stats_output.csv")
```

Figure 5: NLTK based BLEU calculation script example. source: author

The next step involves collecting responses from the selected AI systems and record them in CSV files. These responses are then processed through the script to compute their BLEU scores according to the predefined during scenarios development references. This phase occasionally reveals unexpected results from some systems, which are analyzed later for insights into each system's strengths and weaknesses.

```
weather_inquiry_1,"Today in New York City, it's mostly cloudy with a high of 13°C (55°F) and a low of 7°C (44°F). There's a chance of rain later tonight."
weather_inquiry_1,"New York City is waking up to a cloudy day with a cool breeze. Expect a high of 13°C (55°F) and a low of 7°C (44°F). Keep an umbrella handy, as there's a chance of rain later."
weather_inquiry_1,"Bundle up, New Yorkers! It's a chilly one today with mostly cloudy skies. The high will only reach 13°C (55°F), and the low will dip down to 7°C (44°F). Don't forget your raincoat, as showers are possible tonight."
weather_inquiry_1,"A cool and cloudy day is in store for New York City. Temperatures will hover around 13°C (55°F) with a low of 7°C (44°F). Be prepared for potential rain showers later in the evening."
weather_inquiry_1,"New Yorkers, grab your umbrellas! Today's forecast is mostly cloudy with a chance of rain. The high will be around 13°C (55°F), and the low will drop to 7°C (44°F)."
weather_inquiry_1,"It's a comfortable fall day in New York City with mostly cloudy skies. Expect a high of 13°C (55°F) and a low of 7°C (44°F). There's a slight chance of rain showers this evening."
weather_inquiry_1,"New York City is experiencing a cool and cloudy day. Temperatures will range from 13°C (55°F) at the high to 7°C (44°F) at the low. Be sure to bring an umbrella, as rain is a possibility tonight."
weather_inquiry_1,"Looking out the window in New York City, it's a mostly cloudy day with a cool breeze. The high will be around 13°C (55°F), and the low will dip down to 7°C (44°F). Pack an umbrella if you're heading out, as there's a chance of rain later."
weather_inquiry_1,"New Yorkers, dress in layers today! It's a cool and cloudy day with a high of 13°C (55°F) and a low of 7°C (44°F). Keep an eye on the sky, as there might be some rain showers this evening."
weather_inquiry_1,"Enjoy the crisp fall air in New York City today! The weather is mostly cloudy with a high of 13°C (55°F) and a low of 7°C (44°F). Don't forget your umbrella, as there's a chance of rain showers later on."
weather_inquiry_1,"New York City is experiencing a mild fall day with mostly cloudy skies. The high will be around 13°C (55°F), and the low will dip down to 7°C (44°F). There's a slight chance of rain showers, so be prepared."
weather_inquiry_1,"Looking out the window in New York City, it's a cloudy day with a cool breeze. The high will be around 13°C (55°F), and the low will drop down to 7°C (44°F). Pack an umbrella if you're heading out, as there's a chance of rain later on."
weather_inquiry_1,"New Yorkers, dress in layers today! It's a cool and cloudy day with a high of 13°C (55°F) and a low of 7°C (44°F). Keep an eye on the sky, as there might be some rain showers this evening."
weather_inquiry_1,"It's a fall day in New York City with mostly cloudy skies. Expect a high of 13°C (55°F) and a low of 7°C (44°F). There's a chance of rain, so be sure to bring an umbrella."
```

Figure 6: Gemini weather scenario response csv example. source: author

The data collection process involved querying the AI systems with specific prompts and recording their generated responses for evaluation. To assess the stability of AI systems responses, in addition to conducting BLEU calculations, at the time of data collection, subjective data were also collected for an overall assessment of the success of AI systems responses. Based on them, a DSR analysis was also carried out.

5 Results and Discussion

5.1 General Overview of Testing Results

After receiving all the necessary results and processing the data, we can finally look at the behavior of our systems and discuss the results. These evaluations, based on both single-turn and multi-turn scenarios, offer us insights into the current capabilities and limitations of chosen AI systems in simulating human-like conversations. Before we move on to discussing the calculations, I would like to first focus on each of the systems in order to touch on limitations and anomalies.

- ***Amazon Lex:***

Amazon Lex, while offering robust capabilities for building conversational interfaces, presented challenges during the setup process that could be perceived as user-unfriendly. One of the features of this system is the flexibility on servers to meet the needs of customers, but at the same time this turns into a disadvantage due to the overall complexity of the system. Additionally, its responses tended to be more limited in scope, reflecting a narrower understanding of conversational context. These factors contributed to lower BLEU scores, suggesting that while Lex is a powerful tool for certain applications, it may require further refinement to enhance its responsiveness and ease of setup for broader conversational AI applications.

Unfortunately for us, Amazon cannot offer us such a simple chat agent. The basic settings of the system are not able to cope with even simple questions and creating a chatbot will have such powers can be a serious challenge for an entire development team. In this regard, I exclude Amazon Lex as candidate for comparations due to non-fulfillment of the conditions for the selection of candidates.

- ***Gemini AI:***

Gemini's results are confident, the responses were generally well structured and contextual, resulting in above-average BLEU scores. However, a notable disadvantage was the tendency to receive overly detailed responses. It feels like the responses are being formed through the prism of the search engine, although there is an anomaly in which Gemini cannot respond to the general knowledge request, offering us just to use a Google search for it. Despite this, Gemini's capabilities leave a pleasant impression on me as on user.

- ***ChatGPT:***

ChatGPT as was expected is standout performer in our testing, likely due to its deep learning architecture and extensive training dataset. This AI model demonstrated an impressive ability to generate contextually relevant responses, leading to the highest BLEU scores among the tested systems and the most stable DSR rate. It is very adaptive and can hold the content in very long going chats. The high linguistic quality of GPT's responses highlights its potential as a leading solution in AI-driven dialog systems.

- ***Replika:***

Replika offered the most personalized and human-like interaction experience among the tested systems. Its responses were not only relevant and consistent but also more empathetic and friendly conversational style. This level of personalization and human-like interaction is attributed to Replika's design, which focuses on creating emotionally intelligent dialogues. The quality of interaction with Replica was highly appreciated, however, when it comes to complex queries or tasks, Replica does not give the desired result.

5.2 BLEU Results and DSR Results Overview

After receiving the results of processing our data, I chose a representative way of displaying using graphs. You can view the full summary tables with statistical data in the appendix to this work.

Below are the graphs of the average BLEU score for every chosen system. There are 4 lines on each of the graphs, each of them represents a certain distribution of weights when calculating BLEU (Figure 7). The blue line is a representation of scenarios where preference was given to 1 and 2 grams, so shorter entities were compared with each other for correct linguistics check. The green line represents scenario in which preference was given to longer entities when comparing, so the context of phrases and meaning was compared.

```
weights = [(0.5, 0.5),
           (0.333, 0.333, 0.334),
           (0.25, 0.25, 0.25, 0.25),
           (0.2, 0.2, 0.2, 0.2, 0.2)]
```

Figure 7: Weights distribution for n-grams in BLEU calculation. source: author

Now let’s take a closer look at BLEU scoring result for Gemini. Several visual conclusions can be made from the Gemini graph (Figure 8):

- Observing increased BLEU score for simple weights (blue line), which means Gemini is doing a good job of providing lexically correct information on request.
- Also, the green line shows a very low result, which means that the Gemini system express the general context of the dialogue not so good as at linguistic aspect. Perhaps this is influenced by the search engine legacy of Gemini system.

The general capabilities of Gemini are at fairly high level and I can say that the BLEU score result is relatively low due the specific format of the system's responses, which contains a lot of text.

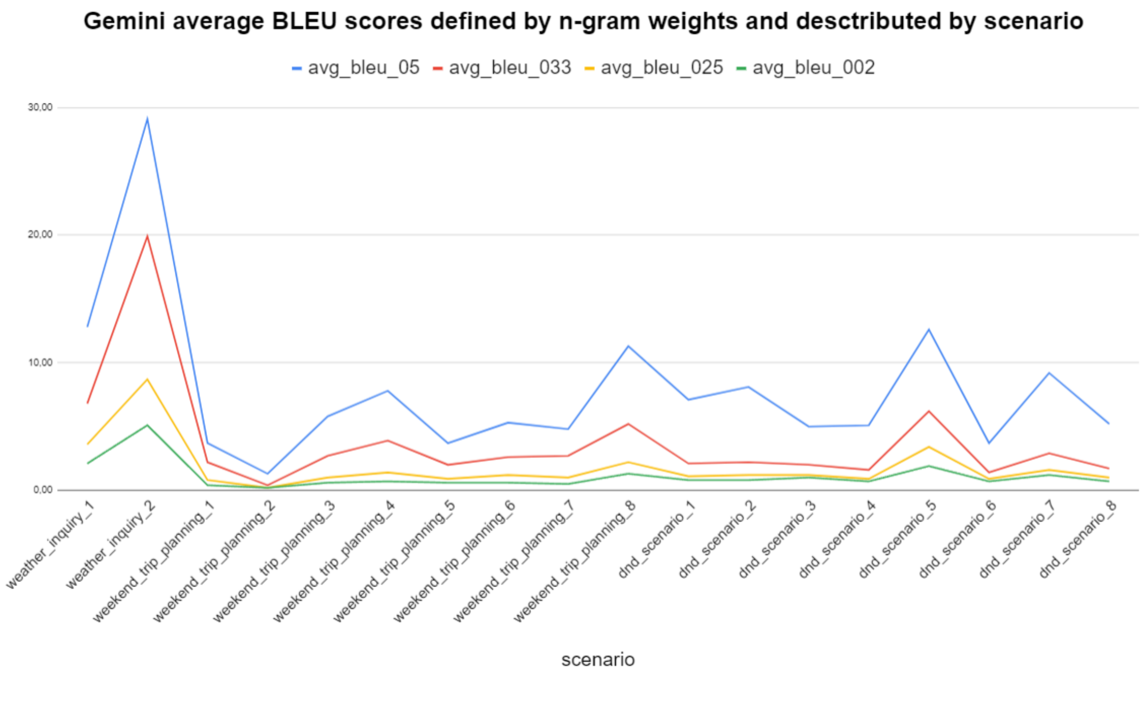

Figure 8:Gemini average BLEU score distribution. source: author

While looking at the ChatGPT graph (Figure 9), I can see it acts as an excellent opposite example of Gemini performance:

- Higher results are immediately visible and therefore the quality of responses is higher than that of the previous system.
- The green line behaves very actively, which means that this system does an excellent job of conveying the context inherent in the combination of words.
- General dynamics of all lines are also clearly observed. They behave at least the same way and have alike patterns, which indicates the comprehensive development and complexity of the ChatGPT system.

This system showed the highest results, but without additional promting it will provide large-volume responses like Gemini, which complicates their processing using BLEU analysis.

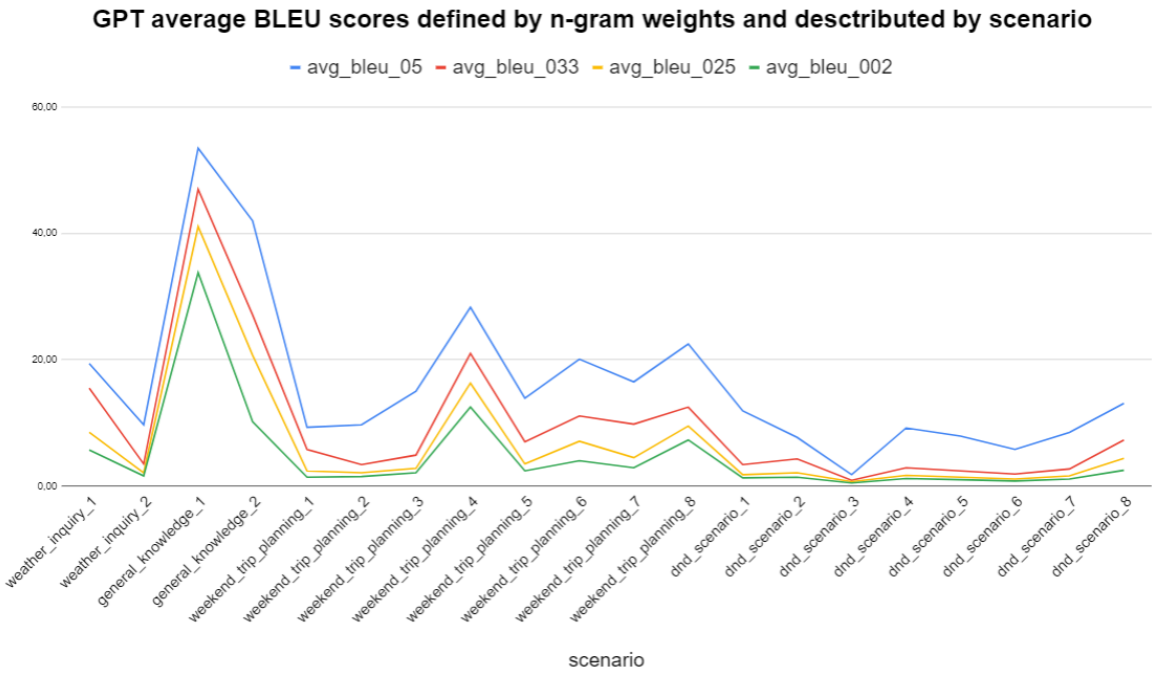

Figure 9: ChatGPT average BLEU score distribution. source: author

As for the Replika graph (Figure 10), the following conclusions can be drawn here:

- The synchronicity of the dynamics of all lines immediately catches the eye. This can serve as a good sign that the system is able to work with both lexical and contextual aspects of the dialogue
- However, the blue line has a fairly large gap between the red one. This may indicate that, although the system is comprehensively developed, it definitely has its strong side - lexically correct answers.
- Also, among other things, it is worth noting the overall BLEU value of the account - it is lower than that of other systems in comparison. This point is expected, as Gemini and GPT have much more resources for development and training.

Replica, unlike other systems, gave the shortest answers, which could have a positive effect on the overall score. However, the quality of these responses was the lowest among the selected systems that passed all stages of testing.

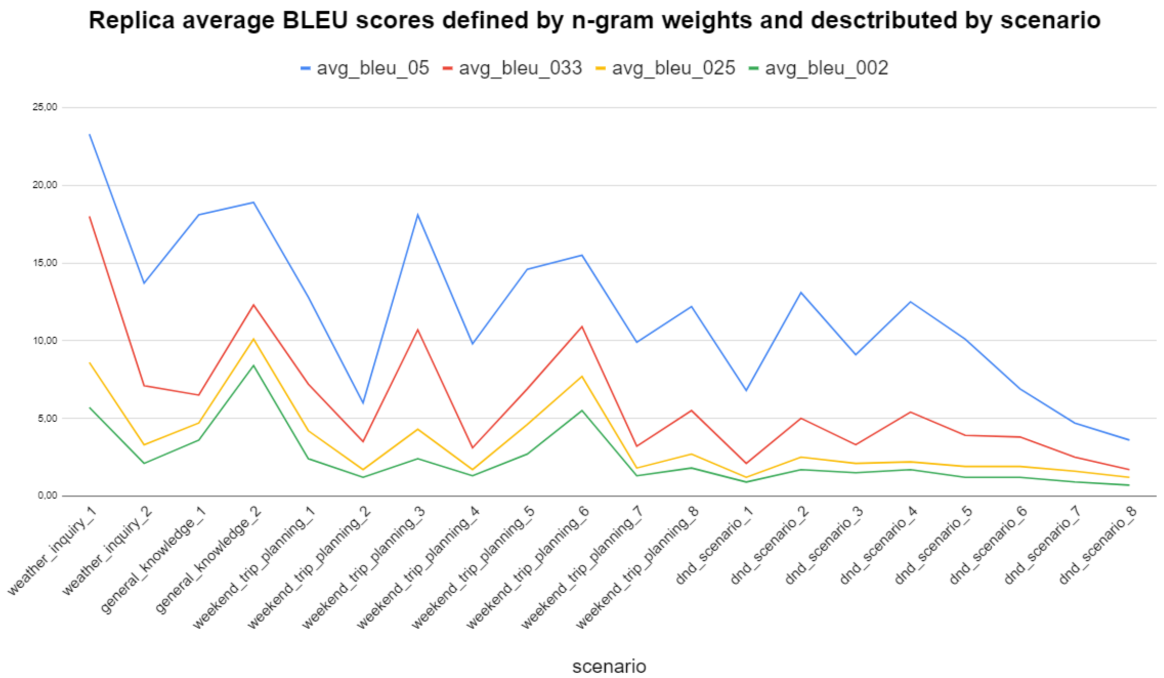

Figure 10: Replika average BLEU score distribution. source: author

In general, thanks to BLEU calculations, it is possible to draw good comparative conclusions about the performance of systems and their limitations. However, this calculation is more suitable for exploring technical capabilities than overall user satisfaction.

In order to find out the test result not from the perspective of text matches, but from the side of task performance, I will use the DSR (Dialog Success Rate) analysis. During the general data collection for BLEU testing, I simultaneously evaluated each response provided by the AI system for conditional success. As justification of success, I took a combination of factors such as the general direction of the dialogue, match with the expected result and the reaction of the system in accordance with the previous context.

According to the results of the DSR analysis, interesting but expected picture appears:

- The ChatGPT system shows a clear advantage in almost every scenario, which once again confirms the development and well-functioning of the system. System provides a more stable result in terms of satisfaction than others.
- On the other hand, the Replica shows the weakest result. Unfortunately, her answers are not always uniform and it's often falls out of the context during the dialogue.
- Gemini anomaly is also clearly visible on DSR, in which system refused to answer questions related to general knowledge and offered to use a search engine instead. However, in another areas, it has shown itself as a fairly strong player in the AI systems market.

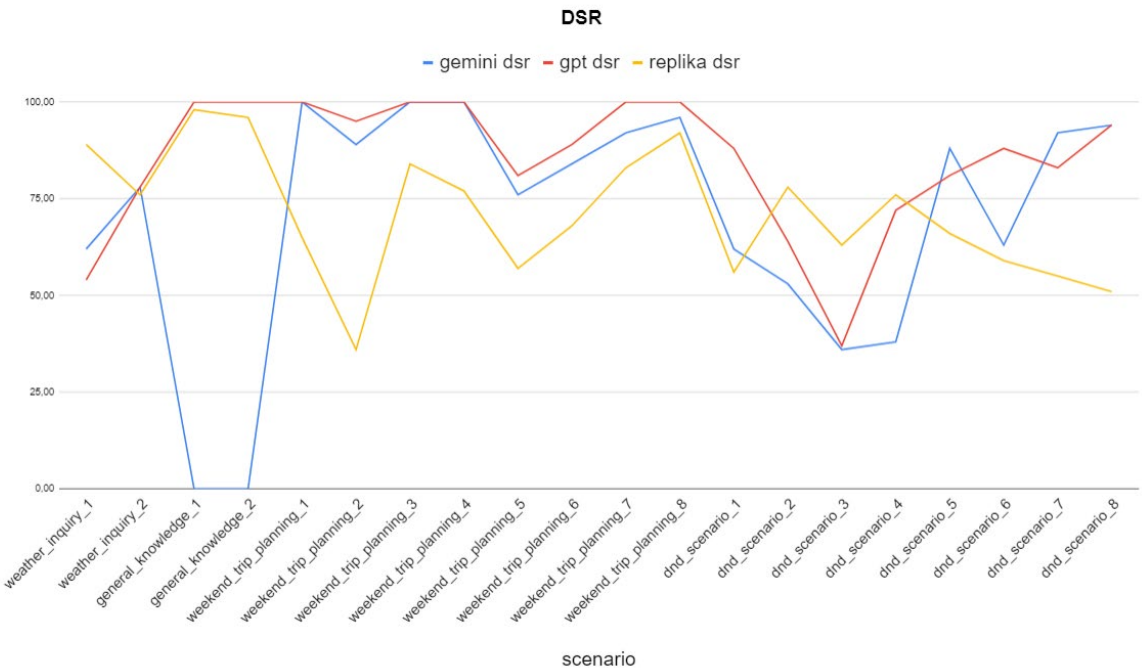

Figure 11: DSR distribution. source: author

6 Conclusion

The objective of this thesis was to evaluate and compare the performance of selected AI dialogue systems in maintaining context consistency, a critical attribute for engaging and effective human-computer interactions. Through careful research, examination of relevant literature and practical testing, this study has shed light on the current state of AI dialogue systems and their ability to understand and retain conversational context over single and multi-turn dialogues.

Through series of carefully crafted testing scenarios, ranging from simple to complex dialogues, this study utilized both BLEU scores and Dialogue Success Rate (DSR) as evaluation metrics to objectively measure each system's performance.

The results highlighted the advanced capabilities of ChatGPT in maintaining context across conversations, demonstrating its superiority in generating coherent and contextually relevant responses. Conversely, the study also identified limitations within other tested systems, such as Gemini AI's tendency for overly detailed responses and Replika's struggle with complex queries, underscoring the diverse challenges faced by AI dialogue systems in achieving context consistency.

In conclusion, this thesis has contributed valuable insights into the current state of modern AI dialogue systems, with a particular focus on their ability to maintain context consistency. It has highlighted problematic areas of AI dialogue systems for future improvement and research. The findings state the importance of continued development in AI technologies to enhance the naturalness and effectiveness of human to AI interactions. As AI continues to evolve, the problem of achievement complete consistency of context will remaining the most important task for AI dialogue systems, determining its future.

7 References

Arora Suket, Kamaljeet Batra, Sarabjit Singh. Dialogue System: A Brief Review. ArXiv (2013). [online] https://arxiv.org/abs/1306.4134; Accessed on October 2023

Akshay Tondak. Amazon Lex Introduction – Conversational AI for Chatbots. 2023. [online] https://k21academy.com/amazon-web-services/aws-ml/amazon-lex/; Accessed on December 2023

Daniel Jurafsky, James Martin. Speech and Language Processing, 2nd Edition. Prentice Hall: 2008. ISBN-13: 978-0131873216

Neri Van Otten. The History of Natural Language Processing & Potential Future Breakthroughs. 2023. [online] https://spotintelligence.com/2023/06/23/history-natural-language-processing/; Accessed on October 2023

Hayet Brabra, Marcos Baez, Boualem Benatallah, Walid Gaaloul, Sara Bouguelia, Shayan Zamanirad. Dialogue management in conversational systems: a review of approaches, challenges and opportunities. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14 (3), pp.783-798. [online] ff10.1109/TCDS.2021.3086565ff. ffhal-03626466f. https://hal.science/hal-03626466; Accessed on October 2023

Kevin Bowden, Shereen Oraby, Amita Misra, JiaQi Wu, Stephanie Lukin, Marilyn Walker. Data-Driven Dialogue Systems for Social Agents. [online] ArXiv abs/1709.03190 (2017):https://arxiv.org/abs/1709.03190; Accessed on October 2023

Soumya Mukherjee. Unveiling Perplexity: Measuring Success of LLMs and Generative AI Models. 2023. [online] https://ramblersm.medium.com/the-significance-of-perplexity-in-evaluating-llms-and-generative-ai-62e290e791bc; Accessed on November 2023

Michael McTear, Zoraida Callejas, David Griol. The Conversational Interface: Talking to Smart Devices. Springer: 2016. ISBN-13: 978-3319329659

Nitin Aggarwal, Amy Liu. KPIs for gen AI: Why measuring your new AI is essential to its success. 2023. [online] https://cloud.google.com/transform/kpis-for-gen-ai-why-measuring-your-new-ai-is-essential-to-its-success; Accessed on December 2023

Patrick Brezillon. Context in Artificial Intelligence: I. A Survey of the Literature. Computers and Artificial Intelligence, Vol. 18, (1999): 321-340. [online] http://www-poleia.lip6.fr/~brezil/Pages2/Publications/CAI1-99.pdf; Accessed on October 2023

Peter Norvig, Stuart Russell. Artificial Intelligence: A Modern Approach, Global Edition 4th Edition. Pearson: 2021. ISBN-13: 978-1292401133

Pierre Lison. A hybrid approach to dialogue management based on probabilistic rules. Computer Speech & Language, Vol. 34, (2015): 232-255. [online] https://www.sciencedirect.com/science/article/pii/S0885230815000029; Accessed November 2023

Sumit Raj. Building Chatbots with Python: Using Natural Language Processing and Machine Learning. Apress: 2018. ISBN-13: 978-1484240953

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, Alexandra Peste. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research, Vol. 22, (2021). [online] https://arxiv.org/abs/2102.00554; Accessed on November 2023

The AI Digital Quill. Deep Inside to Google Gemini: What's Key Features and Why we Use? 2023. [online] https://medium.com/@DigitalQuill.ai/deep-inside-to-google-gemini-whats-key-features-and-why-we-use-ai-gpt-a6576b2e5024; Accessed on February 2024

Tsuta Yuma, Naoki Yoshinaga, Masashi Toyoda. uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems. Annual Meeting of the Association for Computational Linguistics. 199-206 (2020). [online] https://aclanthology.org/2020.acl-srw.27/; Accessed on November 2023

Vijay Kanade. What Is ChatGPT? Characteristics, Uses, and Alternatives. 2023. [online] https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-chatgpt/#_002; Accessed on December 2023

Yang Yang, Juan Cao, Yujun Wen, Pengzhou Zhang. Multiturn dialogue generation by modeling sentence-level and discourse-level contexts. Sci Rep 12, 20349 (2022). [online] https://doi.org/10.1038/s41598-022-24787-1; Accessed on January 2023

8 List of figures

9 List of tables

10 Appendix

scenario	min_bleu_05	max_bleu_05	avg_bleu_05	min_bleu_033	max_bleu_033	avg_bleu_033	min_bleu_025	max_bleu_025	avg_bleu_025
weather_inquiry_1	3,20	20,00	12,80	1,50	13,60	6,80	1,10	9,50	3,60
weather_inquiry_2	19,50	40,70	29,10	11,50	28,00	19,90	5,00	16,30	8,70
weekend_trip_planning_1	2,70	4,80	3,70	1,70	2,90	2,20	0,60	1,00	0,80
weekend_trip_planning_2	0,50	2,70	1,30	0,20	0,70	0,40	0,10	0,40	0,20
weekend_trip_planning_3	5,20	6,60	5,80	2,30	3,20	2,70	0,90	1,20	1,00
weekend_trip_planning_4	6,20	9,90	7,80	3,40	4,20	3,90	1,30	1,50	1,40
weekend_trip_planning_5	1,10	5,40	3,70	0,50	2,90	2,00	0,30	1,20	0,90
weekend_trip_planning_6	5,00	5,80	5,30	2,10	3,60	2,60	0,80	2,10	1,20
weekend_trip_planning_7	4,50	5,40	4,80	2,50	3,00	2,70	0,90	1,10	1,00
weekend_trip_planning_8	10,40	12,90	11,30	3,30	6,70	5,20	1,70	2,50	2,20
dnd_scenario_1	2,00	14,20	7,10	0,90	3,60	2,10	0,60	1,80	1,10
dnd_scenario_2	5,90	10,90	8,10	1,80	2,90	2,20	1,00	1,50	1,20
dnd_scenario_3	2,00	7,30	5,00	1,10	2,60	2,00	0,80	1,50	1,20
dnd_scenario_4	4,70	5,70	5,10	1,50	1,80	1,60	0,90	1,00	0,90
dnd_scenario_5	10,20	14,00	12,60	2,70	8,60	6,20	1,40	5,80	3,40
dnd_scenario_6	1,60	7,10	3,70	0,80	2,30	1,40	0,60	1,30	0,90
dnd_scenario_7	7,30	10,40	9,20	2,30	3,30	2,90	1,30	1,90	1,60
dnd_scenario_8	1,80	8,20	5,20	0,90	2,30	1,70	0,60	1,30	1,00

Appendix 1: Gemini summary statistics part 1. source: author

min_bleu_002	max_bleu_002	avg_bleu_002	avg_resp_char_	avg_scen_char_	avg_incl	dsr
0,90	4,90	2,10	189,80	94,00	31,00	62%
3,00	8,10	5,10	114,10	94,50	30,60	78%
0,40	0,50	0,40	1 884,30	97,50	59,60	100%
0,10	0,30	0,20	1 306,30	100,50	35,30	89%
0,50	0,70	0,60	1 152,30	165,50	48,70	100%
0,70	0,80	0,70	933,30	76,00	82,00	100%
0,30	0,70	0,60	710,70	93,25	50,00	76%
0,40	1,00	0,60	1 620,00	118,25	48,30	84%
0,50	0,60	0,50	1 430,70	105,00	60,00	92%
1,10	1,40	1,30	388,00	76,50	43,80	96%
0,50	1,20	0,80	323,00	103,00	33,30	62%
0,70	1,00	0,80	358,70	95,25	23,10	53%
0,70	1,10	1,00	203,30	67,75	36,40	36%
0,60	0,70	0,70	364,00	88,75	31,00	38%
1,00	2,90	1,90	334,30	87,75	42,90	88%
0,50	0,90	0,70	271,00	88,75	22,20	63%
0,90	1,40	1,20	224,30	82,75	28,20	92%
0,50	0,90	0,70	329,30	94,50	27,40	94%

Appendix 2: Gemini summary statistics part 2. source: author

scenario	min_bleu_05	max_bleu_05	avg_bleu_05	min_bleu_033	max_bleu_033	avg_bleu_033	min_bleu_025	max_bleu_025	avg_bleu_025
weather_inquiry_1	13,50	25,60	19,40	10,70	19,90	15,50	4,50	14,90	8,50
weather_inquiry_2	8,50	11,40	9,70	3,20	4,00	3,50	1,90	2,40	2,10
general_knowledge_1	48,50	56,90	53,50	41,90	54,50	47,00	34,80	52,40	41,10
general_knowledge_2	22,90	59,20	42,00	6,70	42,10	27,10	3,70	32,80	20,70
weekend_trip_planning_1	6,70	10,60	9,30	4,30	7,30	5,80	1,90	2,90	2,40
weekend_trip_planning_2	3,10	15,70	9,70	1,70	4,70	3,40	1,30	2,60	2,10
weekend_trip_planning_3	13,20	17,40	15,00	4,40	5,30	4,90	2,60	3,00	2,80
weekend_trip_planning_4	12,90	47,00	28,30	4,40	36,60	21,00	2,60	27,60	16,30
weekend_trip_planning_5	3,60	28,30	13,90	1,90	14,90	7,00	1,40	6,20	3,50
weekend_trip_planning_6	4,80	34,00	20,10	2,40	24,30	11,10	1,80	15,80	7,10
weekend_trip_planning_7	3,40	23,10	16,50	1,80	14,00	9,80	1,30	6,30	4,50
weekend_trip_planning_8	8,90	34,20	22,50	3,10	27,40	12,50	1,90	22,50	9,50
dnd_scenario_1	9,10	17,40	11,90	2,90	4,30	3,40	1,60	2,20	1,80
dnd_scenario_2	2,40	18,00	7,70	1,20	10,30	4,30	0,90	4,50	2,10
dnd_scenario_3	1,50	1,90	1,80	0,80	1,00	0,90	0,60	0,70	0,70
dnd_scenario_4	8,20	11,00	9,20	2,60	3,40	2,90	1,50	1,90	1,70
dnd_scenario_5	2,00	12,10	7,90	1,00	3,30	2,40	0,70	1,80	1,40
dnd_scenario_6	1,40	14,20	5,80	0,80	3,80	1,90	0,60	2,00	1,10
dnd_scenario_7	5,80	13,40	8,50	2,20	3,80	2,70	1,30	2,10	1,60
dnd_scenario_8	11,60	14,60	13,10	3,40	10,80	7,30	1,90	7,90	4,40

Appendix 3: ChatGPT summary statistics part 1. source: author

min_bleu_002	max_bleu_002	avg_bleu_002	avg_resp_char_	avg_scen_char_	avg_incl	dsr
2,70	11,00	5,70	196,00	94,00	23,80	54%
1,40	1,80	1,60	148,00	94,50	15,70	78%
25,40	50,30	33,80	103,30	65,00	83,30	100%
2,60	15,70	10,20	103,70	93,75	48,90	100%
1,20	1,60	1,40	391,00	97,50	35,30	100%
1,10	1,80	1,50	137,70	100,50	21,60	95%
1,90	2,20	2,10	142,00	165,50	20,50	100%
2,00	20,50	12,50	119,00	76,00	48,70	100%
1,20	3,70	2,40	119,70	93,25	19,00	81%
1,50	7,80	4,00	121,30	118,25	21,70	89%
1,10	3,90	2,90	131,70	105,00	42,20	100%
1,40	17,70	7,30	126,30	76,50	33,30	100%
1,10	1,50	1,30	226,70	103,00	43,70	88%
0,70	2,70	1,40	205,00	95,25	10,30	64%
0,50	0,60	0,50	265,00	67,75	39,40	37%
1,10	1,40	1,20	213,70	88,75	26,20	72%
0,60	1,20	1,00	237,30	87,75	26,20	81%
0,50	1,30	0,80	227,70	88,75	19,40	88%
0,90	1,40	1,10	215,30	82,75	35,90	83%
1,30	4,10	2,50	227,00	94,50	15,70	94%

Appendix 4: ChatGPT summary statistics part 2. source: author

scenario	min_bleu_05	max_bleu_05	avg_bleu_05	min_bleu_033	max_bleu_033	avg_bleu_033	min_bleu_025	max_bleu_025	avg_bleu_025
weather_inquiry_1	17,90	28,70	23,30	13,60	23,80	18,00	7,00	10,70	8,60
weather_inquiry_2	9,20	19,50	13,70	2,70	15,10	7,10	1,50	6,40	3,30
general_knowledge_1	1,80	36,10	18,10	0,80	11,90	6,50	1,80	7,20	4,70
general_knowledge_2	9,60	29,60	18,90	4,00	26,80	12,30	2,70	24,00	10,10
weekend_trip_planning_1	9,60	15,60	12,80	3,40	9,70	7,20	2,00	7,00	4,20
weekend_trip_planning_2	1,80	14,20	6,00	1,00	8,50	3,50	0,70	3,70	1,70
weekend_trip_planning_3	8,30	24,30	18,10	2,60	15,30	10,70	1,50	5,80	4,30
weekend_trip_planning_4	5,90	13,10	9,80	2,00	4,00	3,10	1,20	2,20	1,70
weekend_trip_planning_5	12,30	16,70	14,60	3,50	12,90	6,90	1,90	9,60	4,60
weekend_trip_planning_6	1,60	37,10	15,50	0,80	29,40	10,90	0,60	21,10	7,70
weekend_trip_planning_7	7,60	12,60	9,90	2,30	4,20	3,20	1,30	2,50	1,80
weekend_trip_planning_8	5,90	17,90	12,20	1,90	10,00	5,50	1,10	4,20	2,70
dnd_scenario_1	1,70	13,90	6,80	0,80	4,00	2,10	0,60	2,20	1,20
dnd_scenario_2	11,30	14,90	13,10	3,00	6,80	5,00	1,60	3,20	2,50
dnd_scenario_3	4,30	11,50	9,10	1,60	4,20	3,30	1,00	2,60	2,10
dnd_scenario_4	5,50	29,00	12,50	1,60	20,50	5,40	0,90	13,30	2,20
dnd_scenario_5	1,80	16,80	10,10	0,90	6,30	3,90	0,60	2,60	1,90
dnd_scenario_6	0,90	14,00	6,90	0,50	9,00	3,80	0,40	4,10	1,90
dnd_scenario_7	1,20	2,70	4,70	0,60	1,60	2,50	0,40	1,20	1,60
dnd_scenario_8	1,1	6,40	3,60	0,90	4,70	1,70	0,70	3,60	1,20

Appendix 5: Replica summary statistics part 1. source: author

min_bleu_002	max_bleu_002	avg_bleu_002	avg_resp_char	avg_scen_char	avg_incl	dsr
4,90	6,80	5,70	46,00	94,00	21,40	89%
1,00	3,80	2,10	184,30	94,50	29,40	76%
0,50	5,50	3,60	25,70	65,00	36,70	98%
2,20	20,20	8,40	50,00	93,75	20,00	96%
1,50	3,70	2,40	231,00	97,50	37,30	65%
0,60	2,30	1,20	241,00	100,50	21,50	36%
1,00	3,20	2,40	249,30	165,50	34,60	84%
0,90	1,60	1,30	207,00	76,00	30,80	77%
1,30	5,10	2,70	179,00	93,25	33,30	57%
0,50	15,10	5,50	257,30	118,25	21,70	68%
0,90	1,80	1,30	196,00	105,00	35,50	83%
0,80	2,50	1,80	187,00	76,50	39,60	92%
0,50	1,50	0,90	304,00	103,00	31,20	56%
1,10	2,40	1,70	221,30	95,25	12,80	78%
0,70	1,90	1,50	169,30	67,75	30,30	63%
0,60	6,50	1,70	310,30	88,75	54,80	76%
0,50	1,60	1,20	257,30	87,75	31,00	66%
0,30	2,60	1,20	292,70	88,75	22,20	59%
0,30	1,00	0,90	244,00	82,75	28,20	55%
0,20	2,80	0,70	257,70	94,50	21,50	51%

Appendix 6: Replica summary statistics part 2. source: author