Research Master Thesis

Exploring the Impact of Structured Dialogue Representation on Neural Dialogue Response Generation.

Vasiliki Kyrmanidi

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics

(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Linguistics and Text Mining Lab
Department of Language and Communication
Faculty of Humanities

Supervised by:	Piek Vossen, Lea Krause
2nd reader:	Lucia Donatelli
Submitted:	December 23, 2023

Abstract

In this study, we support the argument that neural dialogue systems relying on structured dialogue representation are more interpretable, controllable and reliable in generating responses. On that account, we seek to explore the impact of structured dialogue representation, formulated as graph triples, on neural response generation focusing on open-domain goal-oriented dialogue. We empirically explore our hypothesis by introducing a series of *qualitative* and *quantitative settings* pertaining to the representation type and amount of dialogue history. Through their combination we formulate distinct configurations of OpenDialKg (Moon et al., 2019) dataset employed to finetune GODEL, a goal directed LLM (Peng et al., 2022). Arguing that structured representation relying solely on factual triples is inadequate for capturing holistically the intricate dialogue properties, we enhance it with perspective triples reflecting dialogue-acts and emotions and investigate their contribution. We comparatively evaluate the impact of our introduced *settings* on the finetuned models' responses by employing standardized automatic NLG metrics and a novel fine-grained manual evaluation framework inspired by the Gricean Maxims. Our analysis reveals preliminary evidence that integrating structured dialogue representation into unstructured dialogue context amplifies model performance. Specifically, structured dialogue history contributes to the models' contextualization, informativeness and reliability, while structured perspectival information enhances naturalness, properties that were found to collectively enhance the overall response quality.

Declaration of Authorship

I, Vasiliki Kyrmanidi, declare that this thesis, titled *Exploring the Impact of Structured Dialogue Representation on Neural Dialogue Response Generation.* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 2024/02/06

Signed:

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisors Prof. Piek Vossen and Ph.D. student Lea Krause for their continuous guidance and reassurance in carrying out this thesis. Thank you for showing me the path into conversational AI, and sharing your enthusiasm for the immersive experience of research. I am deeply thankful to Ph.D. students Selene Baez Santamaria and Annika Kniele, as well as Lea Krause and Prof. Piek Vossen, for their significant contribution to the annotation process and their valuable insights into shaping the annotation guidelines, essential for realizing this project. A heartfelt appreciation goes out to all the members of the Computational Linguistics and Text Mining Lab for fostering the knowledge and skills crucial to undertaking this endeavour, and most importantly for contributing to a transformative and high-quality academic experience here in the Netherlands. Lastly, I want to express special thanks to my partner, Lampis, for his immense emotional support and understanding through the highs and lows of this journey.

List of Figures

Contents

Chapter 1

Introduction

Fifty-seven years following the introduction of of ELIZA (Weizenbaum, 1966) the field of conversational AI has undergone remarkable advancements, with the most recent milestone being the introduction of ChatGPT[1] by OpenAI. The latter has expanded the social implementation of dialogue agents beyond comparison (Abdullah et al., 2022), catalyzing the transition of Natural Language Generation (NLG) research from controlled laboratory settings to real-world applications. This new reality has given rise to a number of ethical considerations concerning the influence and involvement of artificial dialogue agents in society, many of which belonged to the realm of dystopian speculation until recently. While the notion of machines asserting global dominion remains a distant prospect, the evolving integration of conversational agents into various facets of life underscores the imperative to prioritize research on their fair and safe deployment.

State-of-the-art dialogue agents demonstrate a remarkable competence in language generation that closely mirrors human proficiency. Additionally, their training on vast data and integration with external tools armor them with augmented capabilities. These, among else, include access to world knowledge, multilingual capacity, goal-driven behavior, and a semblance of pragmatic understanding reflected through an appropriate use of rhetorical devices and speech acts. Nevertheless, these agents continue to display considerable flaws including the generation of hallucinatory information, weakness in distinguishing between factual and subjective information, and the perpetuation of biases captured in their pretraining. The persistence of such limitations is indicative of a lacking insight into these systems' language learning capabilities, as well as their training data properties and modeling, and calls for the reframing of research directions.

1.1 Motivation

In recent years, Natural Language Generation has witnessed remarkable advancements by developing intricate techniques to formulate responses within a dialogue. Prevailing modern dialogue response generation approaches lean upon statistical deep neural networks that learn probabilistic contextual relations by harnessing extensive datasets. These, also constitute the foundation of pretrained Large Language Models (LLM), such as BERT (Devlin et al., 2018) and GPT-4 (OpenAI, 2023) currently dominating the field with their advanced general-purpose language capabilities. While neural dialogue systems have made significant strides in producing coherent and contextually

[1] https://chat.openai.com/

appropriate dialogue responses, they continue to encounter challenges (Lappin, 2023) and receive critique on their language understanding and learning processes (Marcus et al., 2023).

Traditional language models incorporated explicit human feedback by relying on heuristics and feature processing to make predictions. The transition to neural architectures enhanced flexibility and generalizability in learning input representations, yet human control channels over the input are constrained to input preprocessing and representation techniques. By contrast, transformer-based systems, like LLMs, exhibit autonomy in processing and assimilating information directly from raw input with the only form of external feedback stemming from the selection of their pretraining datasets. To compensate for this, these models rely on prolonged training using extensive computing power and vast corpora of human linguistic expression. However, their deep learning general-purpose-driven architecture renders the monitoring, decoding and modification of their internal linguistic processing mechanisms a formidable task. Placing this issue in the context of dialogue systems, it is challenging to delineate the factors leading to the generation of a response. In addition, the origins of the generated knowledge can be elusive, hindering the distinction between hallucinatory and veracious information, and the tracing of the pathways to hallucination. This deficit in understanding the systems' behaviors, impedes the development of effective counter-strategies and undermines their trustworthiness.

If it is not for their sophisticated processes of language understanding, but rather their massive trainable parameters and data, that neural systems, particularly LLMs, showcase groundbreaking capabilities in dialogue tasks, two critical questions arise: 1) Can we truly assert that this new generation of dialogue systems possesses genuine intelligence, specifically pertaining to effective language understanding and manipulation? 2) If not, why do they consistently achieve high scores in standardized evaluation metrics?. The first question leads us back to the Chinese room argument developed by the American philosopher John Searle (Cole, 2023). It supports that automatic systems' language abilities are merely a result of algorithmic functions of language encoding and decoding, and should not be conflated with the innate cognitive abilities of humans. Framing this argument within the context of neural dialogue systems, endowing them with a form of explicit linguistic knowledge might deem them more self-reliant than operating solely on effective but ultimately naive statistical shortcuts.

Searle also suggests that, since systems are designed to imitate human language abilities, the Turing Test is an inadequate means of evaluation, as it solely focuses on the system's final output. Similarly, conventional word-overlap evaluation methods, have been criticized extensively for their limited depth and insightfulness in evaluating dialogue systems (Srivastava et al., 2023), due to their reference-driven approach. The latter, akin to the Turing Test logic, assesses the quality of predictions by measuring their similarity to human reference responses. Embedding-based approaches follow a similar reference-bound tactic, prioritizing reference-similarity while recent learning-based metrics are susceptible to inherent biases within their own trained probabilistic modeling (Hanna and Bojar, 2021). As such, our second question is addressed—neural dialogue systems attain high performance scores, partly, due to these metrics' fixation on human-likeness rather than language mastery. However, with the influx of novel technologies continuously unveiling more complex system capabilities that the existing techniques fall short in assessing, interpretability becomes a pressing need.

1.2 Relevance

Researchers have undertaken efforts to address the aforementioned challenges in evaluating and interpreting neural-based and, particularly, LLM-driven dialogue systems, that primarily revolve around two key dimensions. Firstly, recent assessment techniques target the models' specific cognitive-like and linguistic capabilities (Gou et al., 2023), as well as their integrated world knowledge (Jiang et al., 2020), yet with a holistic and simultaneously reliable standardized metric remaining elusive. Secondly, in an endeavor to enhance control over the systems' processed input and acquired knowledge, and, thereby, its transparency, tool-augmentation strategies have been proposed. Retrieval Augmented Generation (RAG) is one such approach, integrating a retrieval component to extract externally sourced (structured) knowledge and inject it into the dialogue model (Wu and Zhou, 2021). External structured knowledge can manifest in various forms, including ontologies, knowledge graphs and semantic networks, and can be constructed from diverse sources such as domain-specific repositories, expert-tailored input, or crowdsourced information. Its exploitation to augment dialogue systems has been found to offer a more informative and controlled context for the generation of responses.

Recognizing the potential advantages of structured knowledge in the performance and analysis of dialogue systems, researchers have ventured further by experimenting with structured representations of the primary input per se (Chen et al., 2020a; Wu and Zhou, 2021)—that is the dialogue history without the addition of any external knowledge. In the context of dialogue response generation, in particular, the utilization of structured dialogue history representations in the neural model's training and testing can potentially contribute to the following:

(a) Achieving a more in-depth explicit modeling of the context in the dialogue history by capturing longer and more complex dependencies. This can, in turn, endow the model with enhanced memory reaching back deeper into the dialogue or previous dialogues pertinent to the topic and speaker, thus potentially producing more contextualized responses.

(b) Directing the model's attention to specific information determined by the task and conversational objectives. This may result in enhanced controllability and system adaptability in few-shot settings, potentially, reducing parameterization and data requirements. Additionally, it may also contribute to relevance and the mitigation of generic responses.

(c) Furnishing a more interpretable and transparent framework for tracing and assessing the knowledge embedded in the generated responses. This may ultimately aid in the prevention of hallucinations and enhance system reliability

(d) Capturing better perspectival information encompassed in the dialogue history. This can enhance the model's understanding and expression of perspective, thereby increasing response naturalness and expressivity. Additionally, it can improve the system's ability to distinguish between factual and subjective content.

Despite its promising contributions, the incorporation of structured dialogue input into deep neural models is still in its infancy, primarily, due to the following reasons. Capturing the inherently complex dependencies of a dialogue within a graphical representation is a challenging undertaking. On top of that, the dynamic nature of dialogue, coupled with the substantial data requirements for the adequate training of neural systems, renders the process computationally expensive. Presently, datasets that offer structured dialogue representations are scarce (Moon et al., 2019; Yu et al., 2020; Zhang et al., 2018; Chen et al., 2023), each beset with their own set of limitations. Lastly, transformer-based models, currently dominating the field, are designed to process raw input, hence sidelining research that deviates from this representation, for being counter-intuitive Wu et al. (2023).

1.3 Aim and Research Questions

In this study, we support the argument that neural dialogue systems relying on a structured dialogue representation are more interpretable, controllable and reliable in producing responses. Our primary goal is to explore the impact of structured dialogue history representation on the quality of dialogue responses, and, potentially, contribute to the broader research in developing more sophisticated and effective techniques for improving and evaluating neural dialogue response generation. To achieve our main goal, we comparatively investigate the effect of 7 distinct representation types and amounts of dialogue history on the generated responses, which we define as *qualitative settings* (i.e., relating to the type of dialogue history representation) and *quantitative settings* (i.e., relating to the number of turns used to represent the dialogue history).

Our investigation is driven by the following research question:

> *How does a structured representation of dialogue history impact the quality of neural response generation, as measured by standardized automatic NLG metrics and a manual evaluation inspired by the Gricean Maxims?*

This question can be decomposed into the following sub-questions:

1. How does response quality differ, when the representation of dialogue history is driven by each of the following ***qualitative settings***?
   (a) a ***Structured*** representation in the form of graph triples
   (b) an ***Unstructured*** representation in the form of raw dialogue sequences
   (c) a ***Combined*** representation merging structured and unstructured input
2. How do the following ***quantitative settings*** of dialogue history representation interact with the *qualitative* ones in generating dialogue responses as reflected by response quality?
   (a) ***All*** the past turns
   (b) ***Half*** of the past turns
   (c) ***One*** turn (i.e., the most recent turn)
   (d) ***Shared***: the most recent turn and any prior turns that share at least one common entity with it. The term 'entity' denotes graph entities in structured input and Named Entities in unstructured input.)
3. How does the incorporation of additional perspectival information into the structured representation of the dialogue history affect response quality?

To avoid confusion and facilitate readability the terms displayed in bold will be used consistently in this study as italicized. In addition, to refer to the synergy of a *quantitative* and *qualitative setting* their labels will be concatenated (e.g., *Structured All*)

To address our research questions we generate dialogue responses by employing Godel (Goal-Directed Dialog) (Peng et al., 2022) a transformer-based LLM, designed for knowledge-grounded goal-directed open-domain dialogue. The choice of an LLM-driven model is motivated by LLMs' prominent role in NLG tasks, which creates the imperative to address their aforementioned shortcomings. We finetune Godel on the OpenDialKg dataset (Moon et al., 2019), composed of Q&A-driven dialogues annotated with factual knowledge triples from Freebase (Bollacker et al., 2008). We formulate 11 distinct kinds of input by combining the 3 *qualitative* and 4 *quantitative settings* of dialogue history representation, and produce 11 models. Additionally, arguing that the factual triples constituting the structured representation of dialogues are inadequate for capturing their intricate properties, we enhance structured representations with perspectival information pertaining to dialogue-acts and emotions extracted from every turn, which we also formulate as graph triples. We apply our *settings* to create 7 distinct representations of the enhanced input used to finetune 7 additional models. Finally, we evaluate the quality of model responses through 4 standardized automatic NLG metrics, namely Rouge Lin (2004), Bleu Papineni et al. (2002), Meteor Banerjee and Lavie (2005) and BertScore Zhang et al. (2019). A novel manual evaluation is proposed overhead utilizing criteria inspired by the Gricean Maxims (Grice, 1975)

Before moving further, it is imperative to situate this study within the spectrum of dialogue response generation models spanning between task-oriented dialogue and open-domain chit-chat (see Chapter 2). Merging the features of the two extremes is a desired objective for modern conversational agents, which are expected to fulfill tasks, while carrying an engaging conversation across a range of topics. For instance, an airline customer trying to book a flight online would likely appreciate a friendly chat, additional recommendation or travel inspiration. Similarly, a social chatbot is also expected to perform a range of tasks beyond casual chit-chat. For this reason, this study bridges the gap between open-domain and goal-oriented dialogues, an approach observed in recent works (Joshi et al., 2017).

1.4 Outline

This work is structured as follows: In Chapter 2 we provide a timeline of the dialogue response generation task. We also present an overview of existing dialogue representation approaches along with their respective strengths and weaknesses and discuss briefly graph implementations moving from the broader NLG domain to dialogue systems. Finally, we compare popular dialogue evaluation techniques examining their advantages and limitations.

In Chapter 3 we discuss the requisites for constructing a dataset that incorporates structured dialogue representations and review existing corpora. We present the OpenDialKG dataset, motivate our selection of input representation technique, and outline the preprocessing steps undertaken prior to the training and testing phase. In Chapter 4 we present Godel, and introduce the 18 models distinguished by their input representations and perspective exploitation. In Chapter 5 we present the 4 metrics and 10 criteria employed in our automatic and human evaluation respectively. Additionally, we place the Gricean Maxim theory in the context of dialogue systems and outline existing evaluation techniques it has inspired. We close the Chapter describing the annotation process.

In Chapter 6 we present the automatic and human evaluation results along with their correlation forming initial conclusions on the impact of the distinct representation *settings*. Additionally, we carry out a comprehensive analysis of the disagreement present in human judgments. In Chapter 7 we execute an in-depth error analysis on the instances involved in human evaluation adding to our initial conclusions, while in Chapter 8 we present 3 additional experiments conducted to strengthen our findings.

Finally, in Chapter 9 we summarize our motivation, methodology and key findings. We conclude with a discussion of the principal limitations involved in the study and potential directions for future research.

Chapter 2

Background and Related Work

2.1 Dialogue Systems

Dialogue systems constitute the quintessential means of human-computer interaction bridging computers' capabilities with the human aspect and contributing to a more humane and accessible Artificial Intelligence.

In research, dialogue systems have been classified in a multitude of ways. In this study, we adopt the classification introduced by Young et al. (2022), as it aligns the most with our conceptual understanding of the topic. According to the authors, dialogue systems are distinguished into task-oriented (TOD), such as e-commerce chatbots and open-domain (ODD), such as the recently launched ChatGPT [1]. TODs are designed to fulfill specific user needs, traditionally operating in closed-ended conversations and producing responses confined to predetermined linguistic patterns and topics. By contrast, ODDs typically operate in open-ended conversations and produce domain-agnostic responses resembling those of natural generic chit-chat with the goal of maintaining engagement, and without adhering to structural constraints.

Historically, the distinction between ODDs and TODS was rather straightforward. ODDs relied on statistical data-driven approaches and generic datasets. TODs, on the other hand, relied on monitored development drawing from closed-domain datasets and dialogue state modeling to fulfill predefined dialogue acts (Young et al., 2022). However, new approaches have blurred the lines separating the two types. Efforts to create intelligent and engaging agents with a cross-domain adaptability have combined the knowledge-driven and goal-oriented features of TODs with the open-domain character and chit-chat capabilities of ODDs. Such systems will be the focus of this study and will be referred to as goal-oriented and open-domain.

2.2 Dialogue Response Generation

2.2.1 Definition and Early Approaches

While dialogue systems can be augmented with supplementary features, such as knowledge retrieval, to enhance their effectiveness and usability, their fundamental task remains that of dialogue response generation. As the term implies, dialogue response generation involves the automatic generation of responses to user inputs that are comprehensible to humans, but also contextually relevant to the dialogue history and com-

[1] https://chat.openai.com/

municative objectives. The ultimate goal is for the system to emulate a human interlocutor engaging naturally in conversation. To achieve this, the system must retrieve and process the dialogue context extracting semantic and pragmatic information, based on which, it formulates a response

Early response generation systems were primarily task-oriented. They operated on a pipeline composed of distinct modules namely, a) natural language understanding, b) dialogue state tracking, c) dialogue policy learning and d) natural language generation. The latter was initially realized through conventional rule-based approaches with responses conditioned by predetermined templates, such as tree-structures. Surface realization processes were then used to convert these templates into coherent sequences. Though such systems facilitated control over the topic and conversational style, they produced rather repetitive responses lacking creativity. In addition, they relied extensively on hard-coding, which was highly time and computationally consuming hindering cross-domain adaptability (Chen et al., 2017).

2.2.2 Retrieval vs Generative Models

The rise of data-driven techniques and neural networks improved considerably performance in the task and paved the way towards the development of open-domain dialogue systems. Hard-coded rules were replaced by deep learning algorithms ushering in two possible directions for generating responses, a retrieval-based and a generative approach.

Retrieval-based approaches rely on matching algorithms to identify the most fitting response from a predetermined set of alternatives. Initially, selection was exclusively guided by the dialogue context and confined to single-turn dialogues (Wang et al., 2013). Subsequent approaches incorporated other types of information, such as dependency trees extracted from the past dialogue turns (Wang et al., 2015) or topic vectors sourced from Twitter (Zhao et al., 2011), while the retrieval process soon expanded also to multi-turn dialogues (Lowe et al., 2015).

Despite the success of retrieval-based methods, dialogue response generation has been typically approached as a generative problem, as such techniques tend to yield more natural, creative and engaging responses. Initially, generative and mainly task-oriented dialogue approaches maintained a modular structure (Wen et al., 2015; Zhou et al., 2016) resulting in systems that were prone to error propagation and difficult to update and maintain across all modules (Chen et al., 2017). Such issues were overcome by the adaptation of end-to-end neural approaches on the generative task, which aggregated all modules into a unified framework eliminating the need for hand-crafted feature engineering and revolutionized dialogue systems' performance. The most widely implemented approach is sequence-to-sequence (SEQ2SEQ) learning introduced by Sutskever et al. (2014). It employs an encoder-decoder mechanism relying on neural layers to map an input sequence into an output one. Among the first to implement this approach in dialogue response generation were Vinyals and Le (2015) by deploying the RNN encoder-decoder framework for greedy inference in two-turn dialogues, and Sordoni et al. (2015) by presenting a context-sensitive model that utilized embedding representations of words and phrases in the dialogue along with the RNN decoder architecture of Mikolov et al. (2010). Serban et al. (2017) added to the earlier adaptations by modeling the hierarchical structure inside dialogue sequences, while Xing et al. (2018) expanded the latter's contribution by applying an attention mechanism to the hierarchical RNN model that distinguished between words and whole utterances. However, despite their profound contribution to the task, these models were conditioned only on the recent dialogue history and produced rather generic answers (Li et al., 2015), like *'I don't know'* that are poor dialogue stimulants and lack coherency and consistency (Zhang et al., 2018). To compensate for this, new techniques of encoding context (see Section 2.3) and inducing dialogue specific knowledge were introduced.

2.2.3 Knowledge-Grounded Models

The lack of grounding in the real world is one of the most common obstacles withholding dialogue agents from producing human-like conversation. A knowledge-enhanced response is substantially more informative and engaging, a reality that increased research in knowledge-supported conversational agents.

The combination of external knowledge retrieval with generative approaches has brought dialogue response generation models into a new frontier. Research in this area was stimulated by the prominent public response to agents whose responses are not only natural, but also resourceful, both in the context of goal-oriented and open-domain dialogue. Early approaches often involved incorporating unstructured external knowledge into the response generation process. For instance Long et al. (2017) enhance a task-based response generator with a knowledge extractor that retrieves relevant information from the web, utilizing a CNN knowledge encoder overhead for feature extraction. In their own adaptation of knowledge grounding for open-domain dialogue Ghazvininejad et al. (2018) deploy keyword matching, entity linking and Named Entity Recognition to combine history from Twitter dialogues with textual world facts from Foursquare [2], Wikipedia [3] and Amazon [4]. Similarly, Parthasarathi and Pineau (2018) complement dialogue context with Wikipedia summaries encoded via a Bag-of-Words memory network. Seeking a less sparse representation and, hence, a more scalable implementation, Dinan et al. (2018) combine memory networks with transformers, while amplifying dialogue context with Wikipedia sentences. In contrast to the earlier approaches, the external knowledge accompanying each dialogue is authored explicitly by crowdworkers, facilitating the evaluation of its impact on response quality, a condition also met by the dataset selected for our study.

With the linked data era highlighting the potential of structured knowledge, research in knowledge enhanced dialogue systems changed its direction towards structured knowledge bases. Relevant approaches and their advantages will be presented in Section 2.4.

2.2.4 Large Language Models

As the dialogue response generation timeline unfolds, LLMs emerge as the current state-of-the-art in this domain. Their underlying transformer architecture introduced by Vaswani et al. (2017) incorporates a self attention mechanism that allows for parallel input processing. This enables the capturing of longer dependencies in the input, which can be particularly advantageous within a multi-turn dialogue setting. In contrast to the aforementioned knowledge-enhanced systems relying mainly on controlled querying to retrieve specific information from external resources, LLMs acquire knowledge during their pretraining, on a large-scale and rather autonomous manner. That is, they operate

[2] https://foursquare.com/
[3] https://www.wikipedia.org/
[4] https://www.amazon.com/

on their intricate mechanisms to independently capture the nuances of human language along with world knowledge from vast amounts of data.

Due to their groundbreaking performance in various NLP tasks, research on their application to conversational systems has been extremely dynamic. Open AI's launching of ChatGPT featuring GPT3 and GPT4 [5] is the most recent and trailblazing example. A remarkable performance in dialogue response generation has also been observed by other LLMs typically finetuned on publicly available conversational data. LaMDA (Thoppilan et al., 2022), combined the powers of its large-scale pretraining with supervised finetuning on annotated data and grounding through external knowledge retrieval. BlenderBot 3 (Shuster et al., 2022) demonstrates an enhanced functionality on discrete tasks through modular finetuning and internet access, while it undergoes continuous optimization by publicly interacting with users.

The above models belong in the broader niche of Tool Augmented Language Models (TALM) motivated by the realization that relying fully on increasing scale does not guarantee high-quality-predictions (Parisi et al., 2022). Instead, a synergy with additional mechanisms tailored to overcome specific LLM weaknesses can potentially improve performance. Among these weaknesses, Mehri et al. (2019) identifies their input modeling techniques as inadequately fit for dialogue, given the latter's idiosyncratic features, such as multi-turn dependencies, that pose, among else, advanced coherency challenges. In this work, we follow this theory by experimenting with an alternative form of dialogue representation as a means to augment LLMs' performance.

2.3 Sequential Dialogue Representation

Modern dialogue systems operate in multi-turn dialogue settings, rich in various information types pertaining to speaker profile, dialogue acts, world-knowledge etc and constituting the primary determinants for the generation of responses. It is, therefore, natural that the modeling of dialogue context should receive the necessary attention in conversational AI research.

At its core, dialogue representation follows the principles and development of general text representation. One of the earliest and most popular approaches is that of word embeddings, unique dense vector representations in a high dimensional space capturing the semantic relationships between tokens in the input. Their technology is built upon the distributional hypothesis (Harris, 1954) according to which, the meaning of a word is defined by its context (i.e., surrounding words) and, hence, words that occur in similar contexts are semantically similar. Word embeddings are typically learned via unsupervised representation learning techniques, such as Glove (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013). They have been criticized for their acute sensitivity to training data biases, and, hence, limited generalizability, as well as the limitations of their fixed dimensionality in capturing rich word relations (Sommerauer and Fokkens, 2019).

Sentence embedding approaches were developed in search of a more holistic and informative input representation that goes beyond the limitations of word-to-word relations. Sentence embeddings represent entire sentences or short text extracts as unique high dimensional fixed-size vectors. Popular early techniques include the baseline approach of averaging the word embeddings of a sentence and Doc2Vec (Le and Mikolov,

[5] https://openai.com/

2014). The latter adopts the intuition of WORD2VEC by learning unique vector representations of each sentence in the text, in order to capture document-level semantics. However, both techniques display the same pitfalls as word embeddings, given their context-independent nature. To overcome this issue in the context of dialogue systems, Auguste et al. (2021) introduce Skip-Act-Vectors, a sentence embedding approach that encodes dialogue context into turn embeddings employed in the task of dialogue act prediction. More recently, Zhou et al. (2022) seek to improve sentence embedding representations of dialogue through contrastive learning utilizing consecutive dialogue turns as positive pairs on the assumption that they are semantically related.

Over the years, seeking new methods to exploit the informativeness of dialogue history, researchers have introduced various external mechanisms to explicitly extract valuable dialogue knowledge in a more controlled setting. Zhang et al. (2018) strive to maintain a coherent and stylistically consistent conversation by encoding multi-turn context as GLOVE embeddings and combining it with the speaker's profile information. Additionally, they deploy a key-value memory component to extract relevant information from the dialogue history, drawing inspiration from the memory network architecture proposed by Sukhbaatar et al. (2015). In an effort to improve response relevance without compromising specificity Li et al. (2015) considered the Maximum Mutual Information and Yao et al. (2016) the Inverse Document Frequency between the input dialogue history and the available responses for calculating generative probabilities. By contrast, Choudhary et al. (2017) and Xing et al. (2017) improved the contextual compatibility of the response by introducing topic-aware information extracted from dialogue history.

Nevertheless, contextual encoding was fundamentally improved, not by means of dialogue representation per se, but rather through the introduction of more sophisticated model architectures, with RNNs laying the foundation (see Section 2.2). Though the input representation, typically as word-embeddings, remained unaltered, the hidden states involved in the RNN architecture enabled the filtering and preservation of previously seen information, thus capturing both long and short range dependencies. The latter are typically encoded in the hidden layers of the neural network constituting the so-called contextual embeddings. The most advanced contextual embedding representations are manifested in modern transformer-based LLMs, such as the ones discussed in the previous Section.

In addition to sequential representations, researchers have explored graphical representations of dialogue context. Their different implementations and impact are discussed below.

2.4 Graphs

Before we delve deeper into the utilization of graphs in dialogue systems and dialogue representation, it is important to understand what graphs are and how they have contributed to NLP research.

2.4.1 Definition

Nastase et al. (2015) define a graph as a structured representation of some data. It subsumes a set of vertices (i.e., nodes) $V = v_i \,|\, i = 1, n$, connected by edges $E = (v_i, v_j) \,|\, v_i, v_j \in V$. Nodes represent entities in the data that can be of the same or different type, such as tokens, sentences, documents and concepts. Each node can additionally carry a set of features relevant to the entity they represent. Edges represent semantic, syntactic or frequency relations connecting entities. They may carry a weight, and can be directed or undirected.

Graph structure varies from simple structures of nodes and connecting edges to more complex ones, such as heterogeneous graphs and hypergaphs to account better for the complexity and expressivity of natural language. Determining what the nodes and edges should represent depends on the task and controls the graph informativeness. In other words, any transformation in the topology of the graph can alter the way information is organized and related within the structure impacting the complexity and efficiency of the graph implementation on a given task (Nastase et al., 2015). In the NLP domain, graphs can be categorized into text graphs, semantic graphs, syntactic graphs, knowledge graphs and hybrid graphs depending on the data type they represent (Liu and Wu, 2022).

2.4.2 Application in Natural Language Processing

Traditional graph approaches have found application in several NLP tasks, often in unsupervised or semi-supervised settings. Early algorithms include random walk that was mainly employed for lexical semantic tasks, such as measuring semantic similarity (Ramage et al., 2009) and word-sense disambiguation (Mihalcea, 2005). Graph clustering and graph matching algorithms have also been instrumental in text clustering (Erkan, 2006) and textual entailment (Haghighi et al., 2005) respectively. Label propagation algorithms have been implemented in sentiment analysis (Goldberg and Zhu, 2006), while graph-based data formalisms have proven valuable in exploratory data analysis revealing potentially useful information for executing the task (Nastase et al., 2015).

Despite their extensive applicability in the NLP domain, traditional graph approaches began to subside, as deep learning techniques gained more ground showcasing a promising outperformance. This, was, primarily, attributed to the fact that machine learning algorithms were typically developed for grid-shaped or sequential data, thereby hindering graph-based adaptations from harnessing their powers. In addition, the variability of graph structures is a significant deterrent from developing a unified deep learning framework (Wu et al., 2023).

In response to this gap, Graph Neural Networks (GNNs) emerged after being applied successfully in other domains of AI, such as computer vision and social network analysis, and were accompanied by a surge of interest. GNNs introduced a new graph representation learning framework suitable for various arbitrary graph structures. Each node and edge in a graph representation is associated with a weight (i.e., embedding vector), either learned from scratch or extracted from pretrained models, as in the case of sequential representations. The weight vector of each node is updated in the convolutional layers of the GNN by leveraging information from adjacent nodes to capture better the dependencies in the input. After the graph embeddings are learned, they can be used as input to the same task-specific neural layers of any sequence-based model, eventually, bridging the gap between graphical representations and powerful machine learning algorithms. This integration showcases the potential of GNNs to enhance the representation and understanding of graph-structured information in NLP tasks.

GNNs have been adapted effectively in various NLP tasks, including text classification (Liu et al., 2020), topic modeling (Zhou et al., 2020), sentiment analysis (Chen et al., 2020b), natural language inference (Kapanipathi et al., 2020) and semantic role labelling (Zhang et al., 2020a). Specifically in the area of NLG, where most tasks are translated into a sequence-to-sequence problem, graph-to-sequence (GRAPH2SEQ) models have been developed marking a paradigm shift in addressing graph-related tasks. They combine the benefits of graphical representation in their encoding, with the transformer power in their decoding outperforming several SEQ2SEQ models in tasks, such as machine translation (Yao et al., 2020) and summarization (Zhang et al., 2020b).

2.4.3 Implementation in Response Generation

When it comes to dialogue systems, graphs have been primarily associated with the integration of external knowledge. Such knowledge can be easily circulated due to the linked data paradigm enhancing the interoperability amongst systems and grounding them in the real world. It is typically factual and often domain-specific, and, thus, easier to store in structured databases and retrieve through relatively simple querying. Among relevant research studies, Hixon et al. (2015) translate dialogue corpora rich in factual information into a knowledge graph deployed for question answering. Zhou et al. (2018) enhance the semantic understanding of a user's post with knowledge extracted from a large structured database, to generate responses grounded in commonsense. Similarly, Liu et al. (2021) implement a Graph Convolutional Network (GCN) to reason over the dialogue utterances, while Moon et al. (2019) introduce an attention-based decoding mechanism that walks over the concepts of a knowledge graph, as the dialogue context progresses, and predicts the response concept. Finally, due to their ability to capture complex dependencies in the input, graphs have also been employed for dialogue representation showcasing an encouraging performance, as discussed below.

2.4.4 Input Representation

Having an appropriate graph structure can, potentially, reveal certain patterns in the data that can be otherwise identified through a manual inspection of raw sequences. Graphs have been proved valuable in modeling key-phrases and sentences (Zha, 2002), entities and their relations, such as coreference (Nastase et al., 2015), as well as structured information and coherence at the document-level (Salton et al., 1997).

Typical graph formations, present since early research on the topic, include a) dependency graphs capturing word-level syntactic relations b) constituency graphs capturing phrase-level syntactic relations c) Abstract Meaning Representation (AMR) graphs modeling semantic relations among the input concepts, d) Information Extraction graphs (IE) representing high-level information in the text and revealing knowledge ties amongst sentences, e) Knowledge Graphs (KG) representing world-knowledge relating to concepts and their semantic relations, f) similarity graphs using connecting edges to numerically represent the similarity between node representations.

To enhance the versatility of graph-based representations, but also harness their inherent informativeness, GNN-based learning techniques, such as the aforementioned GRAPH2SEQ can be applied. For instance, Yao et al. (2020) apply GNNs to learn embedding representations from AMR graphs by first converting them into Levi-graphs, while Xu et al. (2018) transform information from an SQL graph into sequential input using the GRAPHSAGE method Hamilton et al. (2017).

In the context of response generation, Hu et al. (2019) employ a directed graph to organize dialogue utterances according to chronological order and speaker. Chen et al. (2020a) compose three distinct graphs capturing token-level, utterance-level and topical similarities. Chen et al. (2020a) combine traditional GNNs with an attention mechanism to acquire semantic network embeddings, while concurrently encoding dialogue utterances, and Hu et al. (2019) deploy a gated GNN to filter the amount of dialogue history considered, when learning utterance-level graph embeddings. These approaches rely on static graphs constructed using the existing prior dialogue knowledge. By contrast, He et al. (2017), Tuan et al. (2019) and Wu and Zhou (2021) tackle the more demanding task of dynamic graph construction by continuously optimizing and updating graphs with new utterances, as the dialogue unfolds. However, either of these works concentrate solely on modeling factual knowledge in the dialogue overlooking other types, such as perspectival information.

This brings us to the realization that graphical input representation is not without its constraints, primarily associated with the graph construction. To begin with, capturing the intricate dependencies between speakers and utterances in the dialogue proves to be challenging. On top of that, in the era of big data, scalability becomes a critical requirement for any automated method, an aspect where traditional graph approaches often fall short computationally, due to their inherent complexity. The latter can grow exponentially, when combined with multi-layer and multi-parameter neural architectures, posing unique challenges to the development of viable models. Additionally, maintaining graphical input can be computationally cumbersome when dealing with dynamic data like dialogue Liu and Wu (2022).

2.5 Evaluating Dialogue Systems

While neural response generation systems have showcased tremendous improvement over the years, the development of more sophisticated technologies necessitates more elaborate evaluation approaches, that are able to capture and interpret the systems' nuanced processes efficiently and systematically. As such, evaluation has emerged as a pivotal step in the development of dialogue systems. However, it presents notable challenges, primarily, due to the non-straightforward nature of the aspects to be evaluated. From a broader perspective, Deriu et al. (2021) identify 5 principal requisites of any evaluation technique: automatization, replicability, interpretability, a strong alignment with manual evaluation and the ability to differentiate between systems. In this Section we review the most prominent evaluation criteria and approaches in the context of dialogue agents. Though, the techniques discussed have been applied across a spectrum of language generation tasks, the focus will remain on textual response generation.

2.5.1 Evaluation Criteria

While there are certain generally considered prerequisites, such as naturalness, grammaticality or relevance, the evaluation of a dialogue response hinges largely on the nature and specific objectives of the system. Traditionally, assessment standards for task-based systems differ from those addressing open-domain ones. Task-success rate and dialogue efficiency are the foremost criteria for assessing task-oriented systems. The former examines the extent to which the user's request is met, while the latter focuses on various factors, such as the number of turns produced until the communicative objective is achieved. In contrast, open-domain systems undergo evaluation through a more flexible framework, due to the absence of a profound objective and the arbitrary structure of the dialogue. Existing approaches concentrate on the overall human-likeness and appropriateness of the generated response or more fine-grained features, such as cohesiveness (Deriu et al., 2021). The evolving convergence of the two types of dialogue agents and their expanded social implementation necessitates a reconsideration of the evaluating criteria and techniques.

2.5.2 Automatic Evaluation Techniques

Automatic evaluation metrics are the principal evaluation approach across language generation tasks presenting both opportunities and challenges. Early automatic approaches involve modeling human judgments into a regression or classification task, to predict user satisfaction (Engelbrecht et al., 2009) or response appropriateness (Lowe et al., 2017). However, the insights resulting from these techniques can be confounded by human annotations and the feature extraction process. The former is susceptible to low inter-annotator agreement compromising the trustworthiness of the process, while latter is highly dependent on dialogue idiosyncrasies delimiting the generalizability of the produced models.

Modern automatic approaches offer, in their majority, a system-agnostic implementation facilitating benchmarking and resulting in a uniform evaluation framework. In addition, they are not grounded in the presence of human annotations and can be distinguished among rule-based, embedding-based and learning based.

The first are conditioned on a set of heuristics and ground-truth responses. Standardized rule-based metrics include Blue (Papineni et al., 2002), Rouge (Lin, 2004) and Meteor (Banerjee and Lavie, 2005). They evaluate the quality of a generated response given the degree of n-gram overlap with its reference counterpart. Yet, such metrics are criticized due to their weak correlation with human judgments (Liu et al., 2016) and their narrow reference-bound scope, which is incompatible with the one-to-many nature of dialogue (Zhao et al., 2017).

Embedding-based metrics, such as BertScore (Zhang et al., 2019) present similar weaknesses. They are reference-dependent, though they attempt to enhance generalizability by focusing on the contextual similarity between the embedding representations of the predicted and reference response.

Compared to early learning-based evaluation approaches (Engelbrecht et al., 2009; Lowe et al., 2017), contemporary metrics in this category are reference free. They model dialogue history and asses its compatibility with the predicted response, typically, through self-supervised training on smaller tasks designed to capture fine-grained dialogue properties. Characteristic examples include USR (Mehri and Eskenazi, 2020b), DIalogRPT (Gao et al., 2020) and FED (Mehri and Eskenazi, 2020a). In their systematic large-scale comparative study, Lan et al. (2020) concluded that learning-based metrics outperform other techniques in terms of capabilities and correlation with human judgments, while their underlying LLM component enhances their generalizability. Finally, their incorporation of smaller-objectives is considered more informative compared to a relatively simplistic comparison with the reference response. Nevertheless, their strong dependency on deep learning architectures renders their findings challenging to deconstruct and interpret.

Finally, in an attempt to understand better the capabilities of dialogue systems and identify areas of improvement, a niche of approaches involve the utilization of adversarial testing. For instance, several approaches propose a checklist of capabilities, which are assessed by observing the dialogue system's performance on perturbated input (Cheng et al., 2019; Zhou et al., 2021). Bruni and Fernandez (2017), on the other hand, utilize a dialogue generator, with the goal of generating human-like output, and a discriminator distinguishing between human and system responses

2.5.3 Manual Evaluation Techniques

To account for the lack of insightfulness observed in automatic evaluation techniques, human judgments serve as an additional means of assessment, as they target qualities carrying the most weight in a specific task. The manual evaluation of dialogue systems is typically conducted via a) lab experiments, where users evaluate their interaction with the system in a controlled setting, b) in-field experiments involving real-user feedback or c) crowdsourcing, where annotators are recruited to evaluate a collection of system responses (Deriu et al., 2021). Despite their effectiveness, human judgments are resource-intensive, challenging to reproduce, and often require specific expertise, thereby limiting their applicability in the dynamic setting of modern dialogue systems.

The above suggests that a holistic and systematic approach is yet to be established, rendering the evaluation of dialogue systems an ongoing challenge. Until such an approach is discovered, researches have been inclined towards the convergence of automatic and manual techniques, recognizing their complementary nature in evaluating dialogue experiments, a line of reasoning also followed in this study.

Chapter 3

Data

3.1 Datasets and Challenges

A core challenge in examining the impact of structured dialogue representation on neural response generation systems is the limited availability of corpora suitable for this purpose. The optimal candidate dataset should exhibit the following key properties contributing to a robust, generalized and holistic experiment.

(a) Each dialogue turn is annotated with a structured representation (e.g., in the form of graph triples).

(b) The structured representation is derived directly from the turn without incorporating any externally-sourced knowledge.

(c) The structured representation differentiates between various information types present in a conversation encompassing both factual and perspectival aspects. Factual information pertains to objective world knowledge, while perspectival information includes the speakers' beliefs, emotions, polarity, sentiment, pragmatic interpretations, perceptions of events, communicative goals etc.

(d) The dialogues cover a diverse array of topics and communicative goals.

(e) The dialogues contain varying numbers of turns of diverse complexity.

At present, there are no datasets satisfying all criteria. Existing corpora that target graph-to-text modeling and accommodate structured dialogue representations come with various limitations. DialogueRE corpus (Yu et al., 2020) is one such example—it consists of open-domain dialogues from the American TV show, 'Friends', annotated with triples of 36 relation types connecting argument pairs, where each triple is linked to a surface trigger expression. Similarly, Harry Potter Dialogue dataset (Chen et al.) consists of dialogues annotated with unique backgrounds encompassing character attributes, geochronological information and speaker relations. Though these corpora reflect the dynamic nature of dialogue through their evolving representations, the domain range of the dialogue topics and the information represented structurally is limited. Most importantly, there is no one-to-one correspondence between graphical representations and dialogue turns. Instead, the annotations are rather dialogue or corpus-based not aligning with our task objectives.

3.2 OpenDialKG

Considering the aforementioned limitations and our study requirements we choose OpenDialKG dataset (Moon et al., 2019) as the most suitable for training and evaluating our models. It is composed of 15K crowdsourced dialogues with a total of 90.5K turns. The dialogues are distinguished between two tasks, recommendation and chit-chat, sharing a common communicative goal of acquiring or exchanging knowledge. This content aligns well with our intention to address a cross-over between task-oriented and open-domain dialogue. While the authors aim to simulate open-domain real-world conversation, the knowledge communicated between the speakers is controlled and centered around 4 topics: 'movies' and 'books' for the recommendation task, 'sports' and 'music' for chit-chat.

To produce a dialogue response, each speaker must select from a predetermined set of factual triples retrieved manually from a filtered knowledge graph from Freebase (Bollacker et al., 2008), a large scale structured database of general factual knowledge that was later merged with Wikidata (Vrandečić and Krötzsch, 2014). Each candidate factual triple is connected to a knowledge entity in the previous turn, via a multi-hop path within the KG. The triples the speaker chooses for generating the response constitute its structured representation and have the form of `[subject, predicate, object]` (see Example (1-a)) or `[object, ∼predicate, subject]` (see Example (1-b)). The resulting dataset was originally employed to train and evaluate a conversational model that performs statistical reasoning by traversing the knowledge graph to find triples relevant to the dialogue context.

Example:

(1) a. `[The Red Violin, has_genre, Mystery]`
b. `[Mystery, ∼has_genre, The Red Violin]`

OpenDialKG addresses adequately conditions (a) and (b) outlined in Section 3.1. However, similar to the datasets discussed above, it is constrained to a subset of the conversational knowledge spectrum that can be graphically represented. This limitation hinders a holistic exploration of the impact of structured dialogue representation and suggests a non-compliance with condition (c). Conditions (d) and (e) are only partially fulfilled. Though the dataset encompasses a small topic variation, these topics are interrelated and centered around a shared communicative objective. Additionally, as Figure 3.1 shows, individual dialogues exhibit varying lengths, yet they generally remain concise with simplistic expressions.

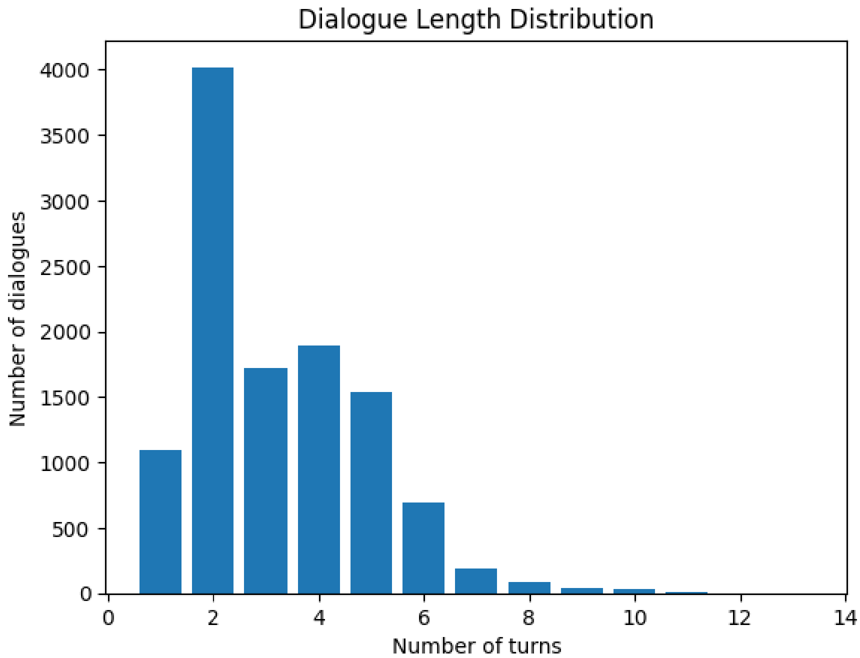

Figure 3.1: Data distribution of dialogue length (i.e., number of turns per dialogue) following the preprocessing steps described in Section 3.4.

Another disruptive aspect pertains to the inaccurate mapping between triples and turns, where certain triples encapsulate information articulated in subsequent turns rather their assigned one. For instance, as illustrated in Example (2), the triples assigned to the second turn are only conveyed at a later point in the dialogue.

Example:

(2) [1]Do you know Holy Hunter?

[2]Sure, the actress and producer. Would you like to know some of her work?
`[Saving Grace, starred_actors, Holly Hunter],`
`[Saving Grace, has_genre, Fantasy]`

[3]Yeah, please. Is she the woman in Saving Grace?

Finally, several turns lack triple annotations, usually, due to one of the following reasons.

(a) They lack factual knowledge. Given that Freebase supports encyclopedic knowledge, only this type of information can be structurally represented.

(b) Their expressed factual knowledge is not sufficient to form a complete knowledge triple. For example, they only express the object entity, while the subject and predicate follow later in the dialogue, as in Example (3), where *'Stephanie Meyer'* is the object of the triple `[Twilight, written_by, Stephanie Meyer]` that has not been fully expressed yet.

(c) Their triples are misplaced, usually to the preceding turn, as in Example (2) above.

Example:

(3) [1]Can you recommend another book by Stephanie Meyer?

3.3 Data Representation Approach

In our approach we frame dialogue response generation as a next-utterance prediction task, where the model considers the preceding dialogue context to make predictions. In our study dialogue history consists of a structured and an unstructured component, therefore we identify two distinct strategies for its representation.

The first strategy involves employing pretrained embeddings to model unstructured history and a graph embedding technique for representing structured history. The embeddings are passed as input to a Seq2Seq model either separately or combined depending on the *qualitative setting* applied. Structured history embeddings can be derived either from pretrained Wikidata graph embeddings, or trained from scratch on a knowledge graph built from OpenDialKG triples employing a GNN-based technique, such as GraphSage Hamilton et al. (2017).

While we consider graph embeddings as the preferred method for capturing structured data relations and exploring their impact, this approach presents formidable challenges. First, representing OpenDialKG triples with pretrained Wikidata embeddings assumes that they are mapped to Wikidata ids. This information is not available in the dataset, potentially, because it was created prior to the integration of Freebase with Wikidata. Secondly, the option of training an embedding model from scratch might also be problematic considering the scarce number, yet relatively diverse content of the triples, as well as their absence in some turns. This can potentially result in an incomplete network with week connectivity and low information transfer.

Besides the above limitations, an embedding representation of dialogue history does not align with the Large Language Model architecture relying on raw textual input processing. Considering the pervasive role of LLMs in contemporary NLP applications, as well as the urgent need of addressing implications associated with them, it appears logical for our work to follow a direction that can be easily applicable in this context. For this reason, we are led to the second strategy for representing dialogue history, which commands a textual representation of both the structured and unstructured context. This approach is comparatively more straightforward to implement and, most importantly, interoperable with the LLM architecture.

3.4 Input Preprocessing

3.4.1 Structured Representation

In order to enhance the number of turns with structured representations we extract an *anchor entity* from the first turn of each dialogue, if it is not alreadt annotated with triples. We define as *anchor entity* a knowledge entity that triggers the conversation and indicates its topic, assuming that the dialogues' communicative goal relies primarily on knowledge exchange and is manifested in the first turn. The *anchor entity* is usually developed into a triple in the second dialogue turn, where knowledge exchange is achieved. Therefore, to extract it, we follow a naive approach of retrieving the triple entity from the second turn, if available, that is also articulated in the first turn. For instance, in Example (4), the triple object, *'Luke Bryan'*, in the second turn is included in the first turn, and therefore, set as the *anchor entity*.

This extraction process, however, does not guarantee that the *anchor entity* will be (accurately) identified. In most cases, this is due to the deprecated structured representation of the second dialogue turn for the reasons described in Section 3.2. We extracted *anchor entities* from approximately 75% of the dialogues and included them in the structured history of their subsequent turns. Though time constrains did not allow for a systematic evaluation of the results, an ad-hoc examination of randomly selected datapoints evidences the efficiency of the process.

Example:

(4) [1]Do you like Luke Bryan, what songs of his are famous?
`[Luke Bryan]`

[2]Yeah I do. I like his song Someone Else Calling You Baby. I just love country music in general. Do you?
`[Someone Else Calling You Baby, Composer, Luke Bryan]`

Furthermore, in order to improve model generalization and pattern learning we normalize the triples in the order of `[subject, predicate, object]` (see Example (5). This results in more meaningful, natural and less syntactically obscure expressions, especially considering the sentence-like textual representation of the triples, described later on in Section 3.4.3.

Example:

(5) a. `[Mystery, ~has_genre, The Red Violin]` [O-P-S]

↓

`[The Red Violin, has_genre, Mystery]` [S-P-O]

b. `[Singer, ~is_a, Selena Gomez]` [O-P-S]

↓

`[Selena Gomez, is_a, Singer]` [S-P-O]

3.4.2 Input Filtering

Taking into account the principal role of structured representation in our study, we opt to exclude turns that lack graph triples from the set of responses used in system training and evaluation. Failing to do so would cause an uneven distribution of data, and potentially compromise the reliability of our conclusions regarding the impact of the distinct *qualitative settings*. This issue would be particularly pronounced in the *Structured setting*, as the absence of triple annotations would deprive completely the response generation model of contextual prompts to rely on.

Similarly, we remove the first turn of each dialogue from the response set, as the absence of preceding context prevents it from being considered a standalone response. Despite not being independent data points, these turns still contribute to the dialogue history of subsequent turns.

Even though the performed filtering reduces the data volume by approximately 60%, we argue that prioritizing quality over quantity secures the transparency and informativeness of our results, yielding more robust conclusions. Our resulting dataset comprises 11,326 dialogues and 36,567 turns, and is partitioned into training and test sets at an 80:20 ratio, determined by the number of dialogues.

Finally, we remove non-essential metadata for our task, such as the number of steps connecting the triples within the Freebase KG and response rating scores. Ultimately, the retained information attributed to each data instance includes a dialogue ID, a turn ID, speaker information (i.e., 'user' or 'assistant'), the raw turn and its structured representation. Utilizing this data, we further extract the turn's unstructured and structured dialogue history corresponding to the preceding raw turns and their structured representations respectively.

3.4.3 Input Formatting

We transform the data according to the format prescribed by GODEL. Every instance is composed by a 'Context' component enclosing the unstructured dialogue history, a 'Knowledge' component carrying the structured dialogue history and a 'Response' component representing the dialogue response. Depending on the *qualitative setting*, 'Context' or 'Knowledge' can be left empty. That is, 'Knowledge' is omitted in the *Unstructured setting* (see Example (6-a)), while the *Structured setting* excludes 'Context' (see Example (6-b)). Finally, the *Combined setting* includes both components (see Example (6-c)).

Example:

(6) a. **Context:** Do you know Holly Hunter? EOS Sure, the actress and producer. Would you like to know some of her work?

Knowledge:

Response: Yeah, please. Is she the woman in Saving Grace?

b. **Context:**

Knowledge: Holly Hunter.Saving Grace starred_actors Holly Hunter. Saving Grace has_genre Fantasy

Response: Yeah, please. Is she the woman in Saving Grace?

c. **Context:** Do you know Holly Hunter? EOS Sure, the actress and producer. Would you like to know some of her work?

Knowledge: Holly Hunter. Saving Grace starred_actors Holly Hunter. Saving Grace has_genre Fantasy

Response: Yeah, please. Is she the woman in Saving Grace?

Besides *qualitative settings*, *quantitative settings* affect the length of ‘Context’ and ‘Knowledge’ by controlling the amount of turns added to the dialogue history. As observed in the examples above, ‘Context’ distinguishes between turns using the special token `‘ EOS ’`. By contrast, ‘Knowledge’ does not support turn distinction—the triples are rather joined together into sentence-like strings separated by a period. Finally, when passed into the model, every instance in the input is compressed into the following format, where the token `<|Knowledge|>` separates the two types of dialogue history and the token `=>` signals the response.

{Context} `<|Knowledge|>` {Knowledge} `=>` {Response}

3.4.4 Perspective Extraction and Representation

As already mentioned, the structured representation of each turn in OpenDialKG does not stretch across the range of conversational knowledge. We judge that we cannot sufficiently draw conclusions on the impact of structured dialogue representation, if perspective information is not structurally expressed. As such, we take our work further by modeling perspective into graph triples.

In particular, we extract the emotions and dialogue-acts from every turn. We apply emotion classification using BERT-BASE finetuned on GoEmotions dataset (Alon and Ko, 2021), which consists of 58K Reddit comments and 28 unique emotion labels. Similarly, we perform dialogue-act classification utilizing a RoBERTa-based multi-label dialogue-act classifier finetuned on MIDAS dataset (Yu and Yu, 2019). The latter is composed by 380K human-machine conversations, where each utterance is annotated with a dialogue-act from 23 distinct categories. The emotion and dialogue-act labels can be found on Appendix A.1 and A.2 respectively.

To represent the extracted perspectival information as graph triples, we define the speaker uttering each turn (i.e., ‘speaker’ or ‘assistant’) as the subject, the perspective type as the predicate, and the corresponding predicted value as the object. An example of a dialogue-act and an emotion triple is provided in Examples (7) and (8) respectively.

Example:

(7) `[user, dialogue act, statement]`

(8) `[assistant, emotion, neutral]`

Finally, the perspective triples of each turn are also converted into sentence strings and appended to the existing factual ones in the ‘Knowledge’ component. Considering the absence of turn distinction, we retain only the perspective triples from the most recent turn in the history. This aims at preventing potential confusion and facilitating the model’s ability to discern meaningful patterns, particularly when confronted with lengthier dialogue histories. In the example below, the perspective triple strings and their corresponding turn are underlined.

Example:

(9) **Context:** Do you know who Suresh Raina is? I'm trying to think where I heard the name from. EOS Suresh Rainia is an athlete on the India national cricket team. Are you a fan of this sport? EOS No wonder I didn't know lol. I am not into Cricket. I know Dwayne Wade tho. EOS Wasn't Dwayne Wade a Point guard in Chicago or somewhere? EOS <u>He is in Miami Heat. Do you like basketball?</u>

Knowledge: Suresh Raina. Suresh Raina is_a Athlete. Dwyane Wade is_a Athlete. Dwyane Wade Position(s) Point guard. Dwyane Wade Position(s) Point guard. <u>user dialogue act statement. user dialogue act yes_no_question. user emotion neutral.</u>

Response: Yes I do. One of my favorites is Lebron James.

Chapter 4

Models

4.1 GODEL

For implementing the dialogue response generation task, we employ Godel (Peng et al., 2022), a generative Large Language Model. The rationale for choosing an LLM lies in its potential to compensate for the limitations found in the dataset's structured dialogue history, as discussed in Chapter 3. Its pretraining can augment the informativeness of the encoded triples, though it also perpetuates inherited biases that are expected to act as a confounding factor in analyzing the impact of the representation *settings* on the predicted responses.

Godel (Grounded Open Dialog Language Model) (Peng et al., 2022) is designed for open-ended goal-directed and open-domain dialogue, properties that are in line with our research goal, chosen dataset and perception of modern dialogue agents. More specifically, Godel supports dialogue grounding on external knowledge in combination with chitchat capabilities, traits that, according to the authors, enhance its utility on a given task, while allowing for an engaging and natural social interaction.

To achieve this duality Godel is pretrained on three distinct types of data. It acquires its general language capabilities from publicly available web text, and its chitchat capabilities from dialogue data sourced from Reddit [1]. For its knowledge grounding, it draws on 4 different datasets designed for knowledge-grounded response generation, task-based dialogue, and conversational question answering. Its multifaceted pretraining facilitates finetuning on various downstream dialogue tasks and domains without requiring large volumes of data. This accommodates our task, considering the substantial data reduction following preprocessing. Godel outperforms other pretrained models, such as T5, BART and DialoGPT, on few-shot datasets according to an automatic and human evaluation conducted by Peng et al. (2022).

The model was originally introduced in three variants, namely GodelB, GodelL and GodelXL, with 220M, 770M and 175B parameters respectively. Given our limited computational resources we choose GodelB featuring a standard Seq2Seq transformer architecture with a 12-layer encoder and a 12-layer decoder. Finally, it is important to emphasize that while the 'Knowledge' component in the model's input was originally designed to incorporate external factual knowledge, in our experiment, it serves the sole purpose of carrying structured dialogue history.

[1] https://www.reddit.com/

4.2 Finetuned Models

We finetune GODEL on 11 distinct configurations of OpenDialKG formulated by combining our *qualitative* and *quantitative settings*, which determine the type and amount of dialogue history respectively. The 11 generated models are described in Figure 4.1.

While the *settings* are self-explanatory in their majority, the *Shared setting* requires further elaboration. It aims to add an additional degree of control in the selection of information employed in response generation. More specifically, it assumes that the dialogue turns, sharing the same conceptual entities with the most recent prompt, include more relevant information for generating contextualized responses and maintaining dialogue coherence. To identify the conceptual entity overlap between the most recent turn and other past turns, we follow a simplistic approach of extracting the verbally expressed graph entities in the case of the *Structured setting* and Named Entities in the case of the *Unstructured setting*. We include in dialogue history only the turns that share at least one common conceptual entity with the most recent prompt. Notably, while the triple entities are inherently included in the dataset, the extraction of Named Entities necessitates further preprocessing. For our purposes, we define a Named Entity as any token-span assigned the label NNP (Proper Noun Singular) or NNPS (Proper Noun Plural) by SPACY's part-of-speech tagger. Given that Named Entities do not necessarily overlap with triple entities, the *Shared setting* is not joined with the *Combined* one.

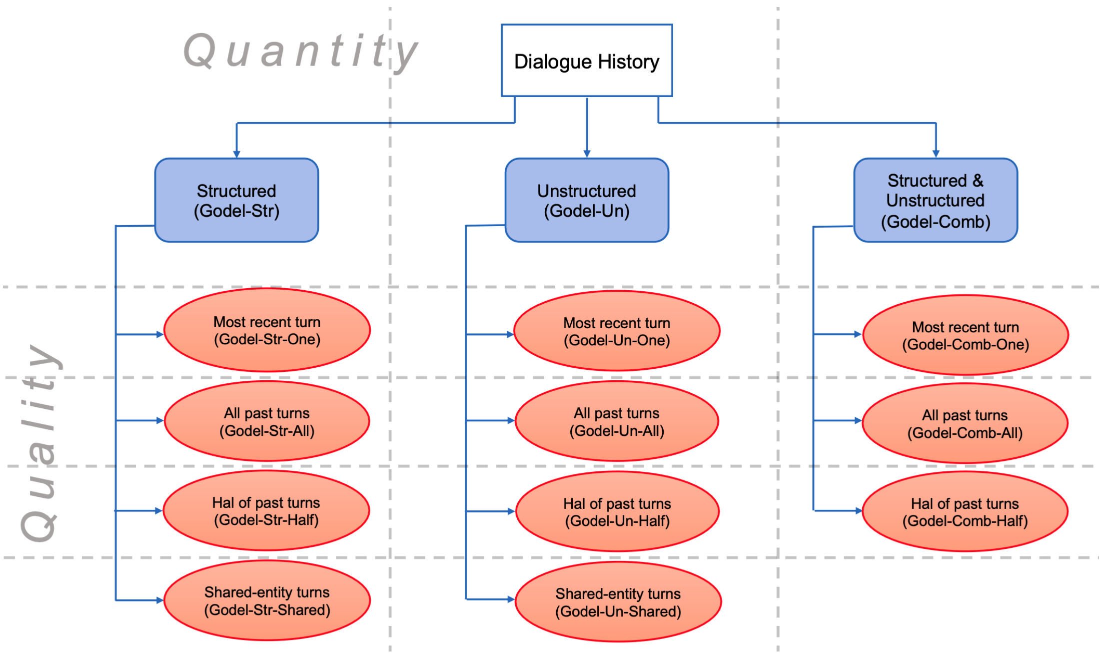

Figure 4.1: The 11 finetuned *non-perspective* GODEL models, categorized according to the *qualitative settings* displayed in blue and the *quantitative settings* displayed in red. The first line of each node explains the setting, while the second line denotes the name of the individual model (eg. GODEL-STR-ONE) or the model group (eg. GODEL-STR) that is finetuned on input formulated according to the corresponding *settings*. These names will be used consistently throughout the remainder of this paper.

To examine the impact of a more holistic structured representation infused with perspectival information, we introduce 7 additional models, described in Figure 4.2). As only the structured representation of the input is modified by incorporating perspective triples, only the *settings* carrying a structured component are applied from the *qualitative* set—that is, the *Combined* and *Structured* ones. To distinguish these models from those exclusively trained on factual information, we will refer to them as *perspective models* in contrast to the *non-perspective models* introduced earlier.

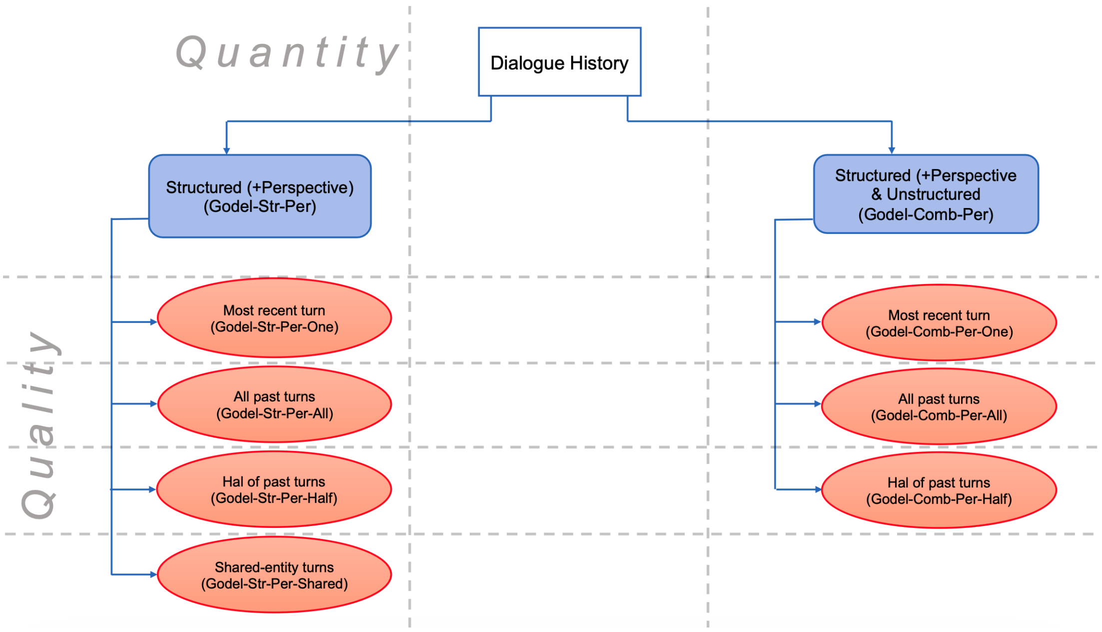

Figure 4.2: The 7 finetuned *perspective* GODEL models, categorized according to the *qualitative settings* displayed in blue and the *quantitative settings* displayed in red. The first line of each node explains the setting, while the second line denotes the name of the individual model (eg. GODEL-STR-PER-ONE) or the model group (eg. GODEL-STR-PER) that is finetuned on input formulated according to the corresponding *setting*. These names will be used consistently throughout the remainder of this paper.

Chapter 5

Evaluation Techniques

5.1 Automatic Evaluation

In conducting the automatic evaluation, we opt for three established word overlap metrics, namely Rouge, Bleu and Meteor, alongside an embedding-based metric, BertScore. This selection is motivated by the fact that these metrics consider diverse aspects in evaluating predictions, thereby offering a more nuanced perspective.

5.1.1 BLEU

Bleu (Bilingual Evaluation Understudy) (Papineni et al., 2002) assesses the precision of the generated text on the basis of its n-gram overlap with one or more reference texts. Precision is computed for each n-gram order (i.e., 1-gram, 2-gram, 3-gram, 4-gram) and the individual scores are, then, combined using geometric averaging. Bleu also incorporates a brevity penalty for predictions that are significantly shorter than the reference. This mitigates the risk of inflated performance scores stemming from the way precision is calculated.

Originally designed for machine translation, Bleu swiftly found application in various language generation subdomains due to its straightforward implementation. In the context of dialogue response generation, it has developed into a standardized metric facilitating benchmarking.

5.1.2 ROUGE

Rouge (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) relies also on n-gram overlap to calculate precision and recall. Its emphasis on the latter, indicating the ratio of reference response n-grams present in the prediction, complements Bleu's precision, which gauges the ratio of prediction n-grams present in the reference response. This renders the combination of both metrics a solid starting point for the evaluation process.

Rouge was first introduced in the domain of text summarization and machine translation, but similarly to Bleu, it has been adapted as a standardized metric for evaluating dialogue generation. It is decomposed into 5 distinct score categories each considering different aspects for measuring f1-score. Rouge-N is based on unigram and bigram overlap, Rouge-L relies on the longest common subsequence between the predicted and reference text, while Rouge-LSum calculates the quotient of the longest common subsequence divided by the sum of the lengths of the generated and reference text. ROUGE-W assigns higher weights to longer shared subsequences, whereas ROUGE-S and ROUGE-SU rely on skip-grams with the latter incorporating a combination of skip-grams and unigrams.

In this study we concentrate on ROUGE-L for two main reasons. Firstly, it is the most widely employed ROUGE performance score in dialogue generation research, facilitating the comparative analysis of our results. Secondly, its focus on the longest common subsequence addresses a notable limitation in our task and dataset. That is, during the creation of OpenDialKG, annotators are presented with a selection of externally-sourced triples for generating a response, with the chosen triple serving as its structured representation. However, in our case, the model operates independently without any external guidance besides dialogue history. As such, employing a word-overlap metric is expected to yield lower scores, as tokens expressing the pregiven external knowledge in the reference are bound to lack in the predictions. On that account, focusing on the longest common sequence appears the most sensible option.

5.1.3 METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) shares common characteristics with the aforementioned metrics, relying on word-overlap and originating from the field of machine translation.

Its distinguishing feature lies on its consideration of explicit word order and various linguistic aspects. Specifically, it relies on unigram-matching between the generated and reference response, which is measured using three distinct mapping modules. The latter focus on diverse scopes of lexical similarity, unlike the one-sided approach of raw token overlap in ROUGE and BLEU. The 'exact' module measures the surface-form overlap between predicted and reference unigrams. The 'porter stem' module measures morphological similarity utilizing PORTER STEMMER (Porter, 1980) to identify root overlap between unigrams. Finally, the 'WN synonymy' module gauges semantic similarity between unigrams using WORDNET (Miller, 1995). The final matching score combines unigram precision and recall with an additional parameter that measures the syntactic appropriateness of the overlapping unigrams, thus also contributing to the assessment of coherence and fluency.

5.1.4 BERTScore

BERTSCORE (Zhang et al., 2019) is an embedding-based metric initially applied on machine translation models. Representing a paradigm shift from traditional metrics that rely on n-gram overlap and independent semantics, BERTSCORE measures the similarity between contextualized embedding representations provided by BERT (Devlin et al., 2018). This approach, mitigates the penalization of semantic and syntactic paraphrasing, observed in the other metrics.

Specifically, BERTSCORE sums up the cosine similarities between the individual tokens in the reference and predicted responses. This calculation enables the capturing of longer dependencies in the text without being bound to strict word order overlap.

BERTSCORE demonstrates a stronger correlation to human judgments and an increased robustness to adversarial data, compared to n-gram overlap metrics, according to a series of machine translation experiments conducted by Zhang et al. (2019). However, a notable drawback besides being reference-bound, is that it has inherited the data biases captured during BERT's pretraining.

The above automatic metric scores range between 0 and 1, where 1 indicates perfect similarity to the reference response. However, in reporting our results, we adhere to the more conventional rescaling of 0 to 100.

In addition, to facilitate n-gram matching we normalize the predicted and reference responses following Peng et al. (2022). This involves lowercasing and removing punctuation, articles and redundant whitespaces.

5.2 Manual Evaluation

The above automatic metrics are primarily designed for application in machine translation and summarization, while being constrained by their reference-oriented and word-overlap approach. It is, therefore, expected that they fall short in adequately covering the intricate requirements of dialogue, as well as its response variability that allows for multiple appropriate responses. To account for such limitations, we introduce a manual evaluation overhead, leveraging criteria inspired by Grice's maxims (Grice, 1975). This approach dissects the assessment of responses into fine-grained dimensions aiming at enhancing the interpretability of the strengths and weaknesses associated with our representation *settings*.

5.2.1 Gricean Maxims

The Gricean Maxims are introduced by the philosopher Grice (1975) in his seminal work 'Logic and Conversation' standing as a pivotal contribution to linguistic and pragmatic research. They are rooted in the Cooperative Principle of communication, which Grice defines as the mutual assumption among interlocutors of a shared effort to communicate meaning efficiently. Within this framework, conversational partners hold 9 specific expectations, Maxims, organized into 4 categories. The first three Maxims primarily concern the content of an utterance, while the latter focuses on the manner in which the conversation unfolds. Grice (1975) describes them as follows:

1. *Quality: Try to make your contribution one that is true.*
   (a) *Do not say what you believe to be false.*
   (b) *Do not say that for which you lack adequate evidence.*
2. *Quantity*
   (a) *Make your contribution as informative as is required*
   (b) *Do not make your contribution more informative than is required.*
3. *Relation: Be relevant.*
4. *Manner: Be perspicuous.*
   (a) *Avoid obscurity of expression.*
   (b) *Avoid ambiguity.*
   (c) *Be brief (avoid unnecessary prolixity).*
   (d) *Be orderly.*

To our knowledge, Saygin and Cicekli (2002) were among the first to apply Gricean Maxims as metrics of human likeness in the evaluation of automatic response generation. Their findings reveal that Quantity and Relation emerge as efficient indicators of system performance, since systems that violate them frequently, are evaluated as unnatural by human interlocutors. Violations of Relation are attributed to either inefficient input processing by the system or a lack of external knowledge. Intriguingly, the violation of Manner is associated with the production of highly human-like responses. Chakrabarti and Luger (2013) follow a similar path assessing the responses of service chatbots against each Gricean Maxim on a Pass-No Pass basis, and concluding that Grice's Maxims constitute an efficient evaluation benchmark for dialogue responses.

In more recent works, Jwalapuram (2017) evaluate a diverse array of systems, spanning from older to state-of-the-art models, across task-oriented and chit-chat scenarios. They observe that the individual Maxim scores correlate with their sum suggesting that they are collectively indicative of a system's overall performance. Khayrallah and Sedoc (2020) introduce the Relative Utterance Quantity (RUQ) metric drawing on the Quantity Maxim, with the aim of quantifying and mitigating *'I don't know'* responses. Finally, Ge et al. (2022) evaluate the quality of generated follow-up questions in conversational surveys using 5 Gricean-Maxim-inspired reference-free automatic metrics, namely Relevance, Informativeness, Truthfulness, Clarity and Coherence, which rely on language models, such as BERT or rule-based and algorithmic heuristics.

5.2.2 Criteria

The above studies demonstrate the relevance and adaptability of Gricean Maxims in the evaluation of dialogue response generation. Following this paradigm, we map Grice's Maxims into 7 criteria that introduce an additional layer of granularity compared to earlier approaches. By increasing specificity, we seek to mitigate the inherent variability stemming from human subjectivity, without disregarding its usefulness, and provide more ground for interpretability. Our criteria are outlined as follows:

1. **Accuracy**: The response should align with the truthfulness and evidentiality aspects of the Quality Maxim, providing conceptually sound information based on common sense and factual knowledge.

2. **Conciseness**: Following directive (2.a) of the Quantity Maxim, the response should not provide more content than necessary for the communicative goal to be addressed and its meaning to be conveyed.

3. **Completeness**: Aligning with directive (2.b) of the Quantity Maxim, the response should provide all the information necessary to effectively address the communicative goal and convey its intended meaning.

4. **Relevance**: In accordance with the Relation Maxim, the response should establish a connection with both the ongoing conversation history and the perceived communicative goal.

5. **Clarity**: As the Maxim of Manner dictates, the response should be free of syntacticosemantic obscurities and ambiguities.

6. **Brevity**: Upholding the same Maxim, the response should avoid unnecessary verbalizations.

7. **Coherence**: Adhering to the final directive (4.d) of the Manner Maxim, the information presented in the response should be semantically, syntactically and logically connected with the recent dialogue history and internally.

To avoid potential confusion, we clarify the distinction between Conciseness and Brevity, with the former pertaining to the conceptual content of an utterance, while the latter to its verbal expression.

We judge that Relevance, Conciseness and Completeness are generally the most challenging to assess, due to the absence of clearly defined boundaries for their violation. For instance, diverging from the conversation topic might seem disruptive, yet intentional deviations can serve to sustain engagement or tactfully redirect discussion. Similarly, despite a response lacking informativeness, it can still consider the communicative goal without necessarily achieving it. For instance, *'I don't know'* addresses the interlocutor's information request by declaring a lack of knowledge or a reluctance to share it. In essence, whether a response violates these criteria hinges on the annotator's perception of the communicative goal and the required information.

Such instances where a Maxim is deliberately breached are defined as 'flouting,' during which speakers intentionally employ conversational implicatures, such as irony and humor, to achieve their communicative objectives (Grice, 1975). The effectiveness of flouting a Maxim without derailing the conversation relies primarily on contextual factors, including the interlocutors' body language, the immediate environment, and shared knowledge—the latter is the only identifiable factor in our task. While flouting is a common phenomenon in human dialogues, we do not anticipate such occurrences in the predicted responses, due to the simplistic nature of our dialogues.

Alongside the criteria outlined above, we introduce 3 additional categories to assess the impact of adding perspectival information on the structured representation of dialogue history. It is important to note, that, considering their broad spectrum, their purpose is primarily exploratory rather than making robust conclusions.

8. **Dialogue-Act**: The response should exhibit an appropriate dialogue-act in line with those in the dialogue history.
9. **Emotion**: The response should display an appropriate emotion given the emotional context of the dialogue history.
10. **Communicative Goal**: The response should successfully meet the communicative goal, expressed in the dialogue history.

5.2.3 Annotation Process

Given our resource and time constraints, we center human evaluation on comparing the distinct *qualitative settings* and discerning the added value of structured perspectival information. We deliberately omit the manual comparison of the *quantitative settings*, as we judge that their minimal impact on model performance, indicated by the automatic metrics in Chapter 6, might be challenging to compare through a few-shot manual evaluation.

Hence, we choose to evaluate the predictions of the following 5 models alongside their respective reference responses. Specifically, we select the best-performing *non-perspective* model within each *qualitative setting* as designated by the automatic evaluation (see Chapter 6)—that is GODEL-COMB-HALF for the *Combined setting*, GODEL-STR-SHARED for the *Structured setting* and GODEL-UN-HALF for the *Unstructured* *setting*. Additionally, we include their *perspective model* counterparts, Godel-Comb-Per-Half and Godel-Str-Per-Shared.

Our annotation set comprises 300 data points, mapped to 50 unique response IDs that each yield 5 model responses and 1 reference response each. To adequately represent the distribution of structured dialogue history lengths in the test data, shown in Figure 5.1, we employ stratified sampling. In particular, we randomly select 10 response IDs with 2 triples in their structured dialogue history, 10 ids with 3 triples and 30 ids with more than 3 triples. Despite their high frequency in the test data, we opt not to include response ids with only 1 triple in their dialogue history. We argue that their comparatively low informativeness may not accurately represent the actual impact of structured representation, especially considering the modest size of our annotation set.

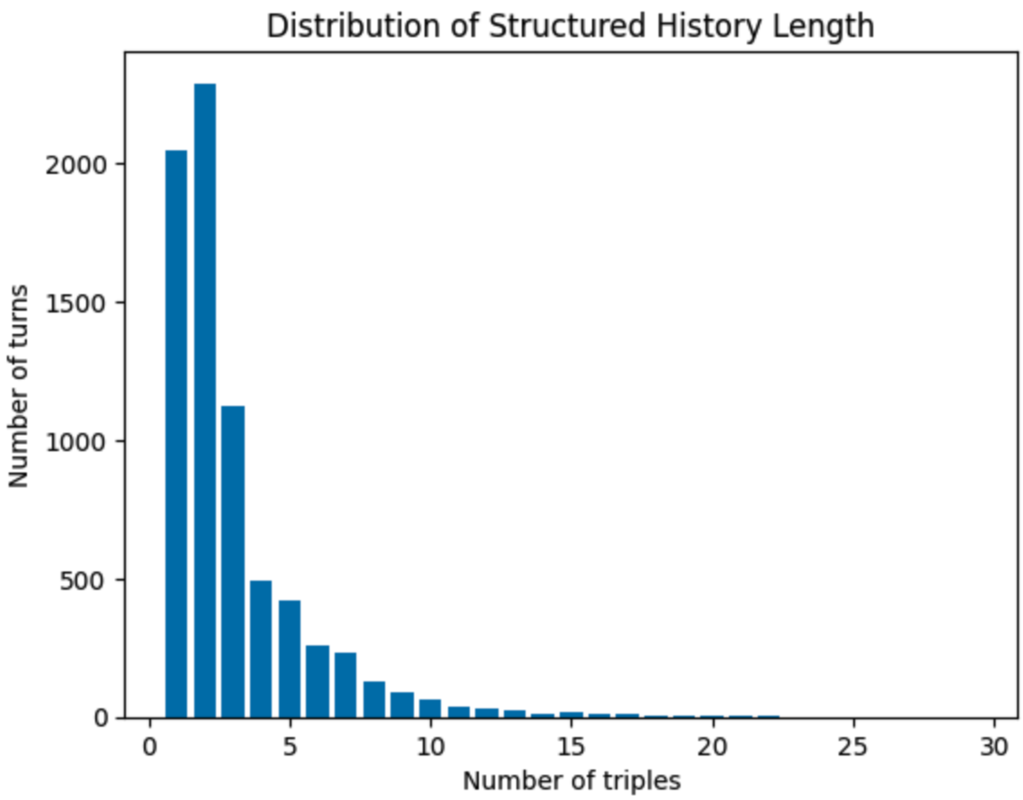

Figure 5.1: The distribution of structured history length in the test dataset, determined by the number of triples representing the dialogue history of each turn, and not the number of turns per se.

We conduct two annotation rounds, where 5 annotators with academic linguistic knowledge form 10 pairs, each annotating 5 response IDs. We assess the first 7 criteria on a Likert scale (1-5), while for the remaining 3, we use the labels Y (Yes), P (Part) and N (No). These labels are later translated into numerical values ranging from 0 to 2, with 0,1 and 2 displaying a positive, neutral and negative performance respectively. More information on the criteria and the annotation guidelines is provided on Appendix B. The annotation results are presented in Section 6.2 of the next Chapter. Finally, to address potential inconveniences caused by the use of Excel formats, such as in visualizing dialogue history, we tailor the annotation tool, introduced by (Camara et al., 2023), and built on the Jupyter Notebook interface.

5.2.4 Disagreement Investigation and Filtering

The original annotation ratings yield a notably low inter-annotator agreement (see Figure 5.3), for which we identify 3 potential determining factors:

(a) A lacking transparency in the annotation guidelines or accountability for the data idiosyncrasies allowing more room for subjective interpretations.

(b) Annotator errors, whether accidental or due to a misinterpretation of the guidelines.

(c) In cases where the first two reasons are not applicable, human subjectivity remains a natural cause of disagreement, hence not targeted in this Section.

Given the pivotal role of human evaluation for the task, and the novelty of its implementation, it is imperative to delve into these factors in an effort to provide insights into the applicability and robustness of the process and its conclusions. The investigation predominantly focuses on our 7 primary Gricean-Maxim-inspired criteria, since, as mentioned earlier, the remaining 3 criteria serve a preliminary exploratory role.

We characterize as outliers those score pairs that exhibit a minimum difference of 3 degrees (e.g., 5-1). We identify a range of 20 to 30 outliers in Soundness, Completeness, Relevance, Clarity and Coherence individually, and approximately 10 in Conciseness and Brevity, which we present in detail in Figure C.1 of Appendix C. Upon examining the most frequently occurring outlier pairs within each category, we notice a consistent difference of 3 degrees in most cases. We argue that such discrepancy is not to be anticipated assuming an adherence to the annotation guidelines, therefore we proceed to a detailed analysis of those instances on Appendix C.

Our findings indicate that human error is the most common cause of disagreement within the outlier pairs, with Relevance, Clarity, Brevity, and Coherence being particularly susceptible. Subjective interpretations emerge as the second most common disruptor of agreement, particularly for Completeness and Conciseness, primarily due to annotators' differing perceptions of the dialogues' communicative goal, while Soundness is influenced by both factors. Our analysis suggests that improving specificity and variety in the guidelines can increase control in interpretation.

Notably, for each criterion, outlier scores are consistently associated with a few specific response IDs, where the predicted responses are often similar or even identical across models, as well as with specific annotators. Besides the examined individual instances, a consistent discrepancy among annotators' rating behaviors is observed in the 'clean' annotation set (see Figure 5.2). These recurrent patterns denote that disagreement can be pinpointed to specific stimuli, and, hence, mitigated, if these are addressed. One potential solution involves conducting trial annotations to provide insights into areas requiring refinement in the guidelines and foster a shared understanding among annotators.

Annotator	Soundness	Conciseness	Completeness	Relevance	Clarity	Brevity	Coherence	Dialogue-act	Emotion	Communicative_goal
a	**4.592**	**4.083**	**3.992**	**3.783**	**4.517**	**4.083**	**3.917**	**0.575**	**0.267**	**0.708**
d	4.617	4.242	4.300	4.300	4.592	4.217	4.450	0.275	0.2	0.625
b	**4.883**	4.467	4.233	4.350	4.892	4.408	**4.517**	0.150	**0.0**	0.625
e	**4.883**	**4.667**	**4.467**	4.108	**4.925**	4.350	4.450	0.108	0.108	**0.342**
c	4.858	4.408	4.308	**4.417**	4.825	**4.617**	4.508	**0.033**	**0.0**	**0.342**

Figure 5.2: Annotators' individual averaged rating behaviors. The values in green and red illustrate the highest and lowest scores assigned for each criterion across annotators.

Though an exhausting investigation of the outliers is impractical given our limited resources, the consistencies observed in our analysis suggest that other cases of disagreement may likely stem from the same factors. As such, deeming outliers as noise, we opt to replace them with the mean score assigned to the entire criterion in the corresponding annotation round. The outlier values are not considered when calculating the mean. By filtering disagreement we produce a 'clean' annotation set that forms the basis for our human evaluation results presented in the next Chapter.

5.2.5 Inter-annotator Agreement

To measure inter-annotator agreement we compute Krippendorff's α coefficient between the 2 annotation rounds formed by vertically concatenating the scores of the 10 annotator pairs.

A cross-comparison between the 'Original Alpha' and 'Clean Alpha' columns in Figure 5.3 reveals that addressing the outlier pairs results in a substantial rise in agreement with values approaching or exceeding the minimum acceptable threshold of 0.67 (Krippendorff, 2004) for most criteria after outlier filtering. Soundness, Coherence, and Completeness show the highest improvement of approximately 0.3 degrees, with the first two achieving an α of 0.7, while the latter displaying almost perfect agreement. In contrast, Clarity and Brevity reach an average agreement slightly below 0.5.

Our 3 exploratory criteria present a considerably lower consensus among annotators. Emotion exhibits the greatest annotation challenge with a negative α score, followed by Dialogue-act that hovers roughly above 0, while communicative goal attains average agreement. These values can be justified by the inherent intricacies involved in interpreting these criteria, such as assessing whether a response conveys a suitable emotional tone and fulfills the communicative goal, or whether there are potentially better-suited alternatives. Considering that the majority of responses comprises statements of neutral emotion and communicative goals that fluctuate between information exchange and chit-chat, annotators' decisions are naturally beset by high subjectivity.

Criteria	Original Alpha	Original MAE	Clean Alpha	Clean MAE	Most Frequent Pairs
Soundness	0.416	**0.507**	0.771	**0.187**	(5-5), (4-5), (5-4), (3-1), (3-4)
Conciseness	0.501	0.550	0.619	0.440	(5-5), (4-4), (5-4), (4-5), (3-3)
Completeness	**0.672**	0.517	**0.913**	0.207	(5-5), (1-1), (4-4), (5-4), (5-3)
Relevance	0.489	**0.727**	0.668	**0.510**	(5-5), (4-4), (5-4), (4-5), (3-5)
Clarity	**0.151**	0.520	**0.485**	0.273	(5-5), (5-4), (4-5), (4-4), (3-5)
Brevity	0.362	0.563	0.492	0.470	(5-5), (4-4), (5-4), (4-5), (4-3)
Coherence	0.378	0.707	0.703	0.403	(5-5), (4-4), (5-4), (4-5), (3-3)
Dialogue-act	0.151	0.350	0.151	0.350	(0-0), (1-0), (0-1), (0-2), (2-0)
Emotion	-0.033	0.223	-0.033	0.223	(0-0), (2-0), (1-0), (0-2), (0-1)
Communicative goal	0.483	0.450	0.483	0.450	(0-0), (2-2), (0-1), (1-0), (0-2)

Figure 5.3: Inter-annotator agreement measured by Krippendorff's α and Mean Absolute Error calculated on the 'clean' annotations following the processing of noisy scores. The highest and lowest scores across criteria are displayed in green and red respectively. The last column displays the 5 most frequently assigned score pairs for each criterion.

Intriguingly, our results reveal that the rise in agreement does not consistently align with the quantity of outliers. This stresses the necessity to dive deeper into the calculation of Krippendorff's α, in order to substantially comprehend how disagreement manifests, and thereby identify the annotation challenges associated with each criterion.

Krippendorff's Alpha Calculation:

$$\alpha = 1 - \frac{O_{ij}}{E_{ij}}$$

- O_{ij} is the observed agreement between raters i and j, calculated as

$$\frac{MSB - MSW}{MSB + (k-1) \cdot MSW}$$

- MSB represents the Mean Squared Error (MSE) between the mean score of round1 and round2.
- MSW represents the MSE between each pair between the annotation rounds.
- E_{ij} is the expected agreement between raters i and j, calculated as

$$\frac{1}{N}\sum_{k=1}^{N}(n_{ik} \cdot n_{jk})$$

- N is the total number of annotation pairs.
- n_{ik} and n_{jk} are the number of times raters i and j assign category k.

The above indicate that Krippendorff's α is highly dependent on the frequency of individual scores occurring separately and in pairs, and, hence, more robust to outliers. For instance, disagreement caused by an infrequently assigned pair (e.g., 5-1) contributes less to α, compared to disagreement concerning a frequently assigned pair (e.g., 5-4). Moreover, the coefficient is sensitive to the magnitude of disagreement, with a quadratic penalty applied. However, these properties also render it difficult to interpret in contexts with high annotator error.

In an attempt to address this challenge and elucidate the influence of score frequencies in calculating α, we also provide the Mean Absolute Error (MAE) per category in Figure 5.3, where a low value, signifies a high overlap between annotation rounds and is solely influenced by the amount of disagreeing pairs and their absolute rating difference, without being affected by the score distribution.

To contextualize our understanding of the above agreement metrics' synergy, we categorize our 7 primary criteria into the following 4 scenarios seeking to capture their agreement and disagreement patterns and offer insights into the intricacies of inter-annotator reliability.

- **High α - low MAE** indicates high agreement on frequently assigned scores, with few disagreement cases likely involving minor score differences. This scenario is observed in Completeness and Soundness.

- **High α - high MAE** indicates high agreement on frequently assigned scores, but also a notable number of disagreement cases, where score differences are likely to be larger. This pattern is prevalent in Relevance, Conciseness, and Coherence.

- **Low α - low MAE** involves low agreement on frequent scores, but with few disagreement cases revolving around small score differences. Clarity belongs to this case.

- **Low α - high MAE** is marked by low agreement on frequent scores and a relatively high number of disagreement cases that tend to involve larger score differences. Such patterns are found in Brevity.

Given the above categorization, we rely on the most frequently assigned score pairs per category in Figure 5.3, and the detailed distribution of scores in Figure C.2 of Appendix C to draw the following arguments. Annotators highly agree that most of the responses score high in Completeness, Soundness, Relevance, Conciseness and Coherence. However, for Completeness and Soundness consensus pervades all score values, while for the remaining three, disagreement often concerns the lower end of the score scale. In the case of Clarity and Brevity, annotators appear to be more divided, with discrepancies pertaining to higher values for Clarity, and to higher against lower values for Brevity.

Chapter 6

Results

6.1 Automatic Evaluation Results

6.1.1 Non-perspective Model Scores

Examining the performance of *non-perspective models* on the automatic metrics, displayed in Figure 6.1, we observe that GODEL-COMB models consistently outperform GODEL-UN models, on a similar level across all metrics. This indicates the advantage of the *Combined setting* over the *Unstructured* one. The *Structured* setting, on the other hand, has the least positive impact, as indicated by the scores achieved by GODEL-STR models, aligning with our expectations that the available triples cannot sufficiently represent the multifaceted nature of dialogue. Interestingly, GODEL-STR scores, particularly in BLEU and BERTSCORE, are only marginally lower than GODEL-UN scores. This potentially suggests that, despite the disparity in input quality, the *Structured setting* is capable of maintaining models' attention towards core content over ancillary conversational details.

Comparing the impact of *quantitative* settings we detect that including half of the dialogue history (*Half Setting*) yields superior results for GODEL-COMB and GODEL-UN models, while GODEL-STR models perform best under the *Shared* and *All settings.* In contrast, while including only the most recent turn (*One setting*) results in the second best performance for GODEL-COMB models, it has an adverse effect on GODEL-UN and GODEL-STR models. This observation implies that the *Combined setting* interacts best with smaller amounts of history, while the *Unstructured* and *Structured settings* require larger input quantities, a distinction discussed further in Chapter 9. Finally, the relatively high scores of GODEL-UN-SHARED and GODEL-STR-SHARED evidence the *Shared setting*'s efficacy.

	Model	ROUGE-L	BLEU	METEOR	BERTSCORE
L	Godel-Comb-All	**21.6600**	**6.0216**	**19.3691**	**75.4350**
T, L	Godel-Comb-Half	**22.3572**	**6.4409**	**20.0572**	**75.7344**
L	Godel-Comb-One	**21.7386**	**6.2513**	**19.4251**	**75.5395**
	Godel-Un-All	20.8947	5.5773	18.7281	75.1405
T	Godel-Un-Half	21.1473	5.7411	18.8452	75.3333
	Godel-Un-One	20.0750	5.3311	17.8797	74.9071
L	Godel-Un-Shared	20.4384	5.4379	18.2197	75.0224
	Godel-Str-All	17.9490	4.9101	15.9652	73.9270
	Godel-Str-Half	17.5627	4.6915	**15.5181**	73.8779
	Godel-Str-One	**17.3009**	**4.6319**	15.6147	**73.7705**
T	Godel-Str-Shared	18.0096	4.8018	15.9229	74.0518

Figure 6.1: ROUGE, BLEU, METEOR and BERTSCORE performance scores of *non-perspective models*. For every metric the highest model scores are displayed in green from darker (1st highest score) to lighter (3rd highest score), while the lowest score is displayed in red. Models labeled with the letter "T" on their left-side perform the best across *quanTitative settings* (e.g., GODEL-COMB-HALF outperforms GODEL-COMB-ONE and GODEL-COMB-ALL). The letter "L" indicates the best performing model across *quaLitative settings* (e.g., GODEL-COMB-HALF outperforms GODEL-UN-HALF and GODEL-STR-HALF).

As part of contextualizing our findings, in Table 6.1, we present the performance of GPT2 and BART finetuned on OpenDialKG (Bang et al., 2023), along with ChatGPT applied on 50 data samples (Ji et al., 2022). These results serve only as reference points of how alternative approaches perform on this data considering variations in data splitting, preprocessing, model parameterization and pretraining data. GPT2 and BARTT surpass largely the performance of our finetuned GODEL models, mainly, due to task differences. Contrary to our implementation, Bang et al. (2023) include the structured representation of the target response in the dialogue history for knowledge grounding, leading naturally to greater content overlap between predicted and reference responses. In contrast, CHATGPT performs comparably to GODEL-STR models, even lower according to BLEU. However, Ji et al. (2022) caution against relying solely on automatic metrics, since ChatGPT's efficacy was more pronounced during human evaluation. This underscores the need to explore whether human judgements unveil different behaviors in our models, from those observed by automatic metrics.

Model	ROUGE-L	BLEU
GPT2	29.5	10.2
BART	33.3	14.4
ChatGPT	18.6	4.1

Table 6.1: Scores of BART, GPT2 (Ji et al., 2022) and ChatGPT (Bang et al., 2023) applied on OpenDialKG

6.1.2 Perspective Model Scores

The scores of Godel-Str-Per models in Figure 6.2 reveal that the incorporation of perspective triples optimizes the performance of Godel-Str models by approximately 1 degree across all metrics. This appears as a logical outcome considering that perspectival information enhances the limited informativeness of the *Structured setting*. Conversely, Godel-Comb models, with the exception of Godel-Comb-All, experience an opposite effect, arguably due to their lower information needs, discussed earlier.

Overall, *perspective models* exhibit similar behavior to their *non-perspective* counterparts. The *Combined setting* yields the best overall results interacting best with the *Half setting*, while the *Structured setting* collaborates best with the *Shared* one. Additionally, the impact of quantity fluctuation on Godel-Str-Per models is again inversely analogous to that observed in Godel-Comb-Per models.

	Model	ROUGE-L	BLEU	METEOR	BERTSCORE
L	Godel-Comb-Per-All	+21.6627	+6.0745	-19.3554	+75.4518
T, L	Godel-Comb-Per-Half	**-22.3446**	**-6.3344**	**-19.8251**	**-75.6293**
L	Godel-Comb-Per-One	-21.4501	-6.1212	-19.2006	-75.3442
	Godel-Str-Per-All	+19.0429	**+5.1898**	+16.7248	+74.5506
	Godel-Str-Per-Half	+18.8557	+5.2710	+16.6936	+74.4334
	Godel-Str-Per-One	**+18.7124**	+5.3118	**+16.3829**	**+74.4323**
T	Godel-Str-Per-Shared	+19.2941	+5.304	+17.0117	74.5835

Figure 6.2: Rouge, Bleu, Meteor and BertScore performance scores of *perspective models*. For every metric the highest model scores are displayed in green while the lowest score is displayed in red. Models labeled with the letter "T" on their left-side perform the best across *quanTitative settings* (e.g., Godel-Comb-Per-Half outperforms Godel-Comb-Per-One and Godel-Comb-Per-All). The letter "L" indicates the best performing model across *quaLitative settings* (e.g., Godel-Comb-Per-Half outperforms Godel-Str-Per-Half) The symbols '+', '-' and '=' preceding the numbers, denote increase, decrease or stability in performance respectively compared to the scores of their corresponding *non-perspective models*.

6.2 Human Evaluation Results

6.2.1 Human Judgement Scores

The reported human evaluation results in Figure 6.3 are extracted from the 'clean' annotation set, created to address potential noise stemming from outlier scores, as discussed in the previous Chapter. They reveal a consensus with automatic metrics in ranking model performance, with the *Combined setting* having the most positive effect overall—GODEL-COMB-HALF is the top-performing model in most aspects, though slightly surpassed by GODEL-UN-HALF in Conciseness and Clarity. In accordance with automatic scores, the *Structured setting* deprecates response quality with GODEL-STR-SHARED being the most negatively rated model, followed by GODEL-STR-PER-SHARED.

GODEL-COMB-PER-HALF ratings denote that the introduction of perspective has a more positive impact on the *Combined setting* than indicated by automatic metrics, for Completeness, Dialogue-act and Communicative goal, yet not for Clarity Relevance, Brevity and Emotion. However, for those criteria where GODEL-COMB-PER-HALF is rated lower compared to its corresponding *non-perspective model*, it also falls behind GODEL-UN-HALF, a trend not observed in the automatic scores. When combined with the *Structured* setting, perspectival information shows to improve response quality throughout, except for Clarity, as discerned by GODEL-STR-PER-SHARED ratings. Finally, while reference responses generally outscore the evaluated models, they vaguely fall behind in Clarity and Emotion.

Overall, while their comparison is feasible, all model ratings surpass the median score of 3, with the majority exceeding 4 and exhibiting close similarity. The latter likely hinges on the models' shared linguistic pretraining stemming from their common underlying LLM, that may lead to the generation of closely similar responses.

Criteria	Soundness	Conciseness	Completeness	Relevance	Clarity	Brevity	Coherence	AVERAGE	Dialogue-act	Emotion	Communicative goal
Reference	**4.920**	**4.540**	**4.880**	**4.720**	**4.780**	**4.430**	**4.690**	**4.709**	**0.110**	**0.110**	**0.160**
Godel-Comb-Half	4.840	4.500	4.580	4.400	4.820	4.420	4.580	4.591	0.150	0.050	0.370
Godel-Un-Half	4.780	4.540	4.560	4.280	4.830	4.360	4.490	4.549	0.130	0.070	0.450
Godel-Str-Shared	4.660	3.920	3.270	3.560	4.630	4.200	3.700	3.991	0.550	0.190	1.040
Godel-Comb-Per-Half	-4.710	-4.460	+4.590	-4.360	=4.820	-4.360	=4.580	-4.554	+0.110	-0.110	+0.360
Godel-Str-Per-Shared	+4.690	+4.280	+3.680	+3.830	-4.620	+4.240	+4.170	+4.216	+0.320	+0.160	+0.790

Figure 6.3: Human evaluation ratings of the selected models and reference responses over 10 criteria. The scores are averaged over the clean version of human annotations adjusted for noise. Reference response scores are displayed in bold. The highest and lowest-scoring models per category are displayed in green and red respectively excluding reference responses. The symbols '+', '-' and '=' preceding the scores of *perspective models* denote increase, decrease or stability in performance respectively compared to the scores of their corresponding *non-perspective model*. The scores of the last three criteria range from 0 to 2, with 0, 1 and 2 signaling a positive, neutral and negative performance respectively

6.2.2 Correlation with Automatic Scores

Finally, we calculate Pearson correlation on response-level between human judgements and automatic metric scores. The latter are obtained from the same 300 responses used in human evaluation. Overall, we observe a rather weak correlation among all automatic metrics and human evaluation criteria, which is consistent with prior studies (Liu et al., 2016; Deriu et al., 2021), and anticipated given their inherently distinct perspectives. Among the automatic metrics, BLEU exhibits a slightly lower correlation with human judgements across all criteria. Completeness, Relevance, and Coherence consistently demonstrate the highest correlations with all automatic metrics that ranges between 0.2 to 0.3, while Dialogue-act and Communicative goal display a negative correlation, albeit not considerably below 0. Interestingly, only the highest and lowest correlation results prove to be statistically significant, suggesting that a more extensive annotation process may be necessary to draw conclusive insights for the remaining criteria, whose correlation with automatic metrics hovers around 0.

Despite the absence of a consistent linear relationship in their individual scores, both human judgments and automatic metrics exhibit a consensus in their overall model ranking, except for GODEL-COMB-PER-HALF. This implies that, while the two evaluation approaches concur on the overall quality differences between the models' responses, they disagree on which responses and properties determine that quality.

Criteria	ROUGE-L	BLEU	METEOR	BERTSCORE
Soundness	0.100	0.102	0.097	0.083
Conciseness	0.116 *	0.109	0.091	0.087
Completeness	**0.275 ***	**0.238 ***	**0.272 ***	**0.250 ***
Relevance	**0.324 ***	**0.247 ***	**0.318 ***	**0.309 ***
Clarity	0.068	0.042	0.065	0.049
Brevity	0.085	0.09	0.056	0.08
Coherence	**0.244 ***	**0.176 ***	**0.239 ***	**0.228 ***
Dialogue-act	-0.18 *	-0.144 *	-0.167 *	-0.16 *
Emotion	-0.014	0.01	-0.005	0.032
Communicative goal	-0.322 *	-0.267 *	-0.297 *	-0.274 *

Figure 6.4: Pearson correlation between automatic scores and human judgments on the 300 instances included in human evaluation. The 3 highest correlation scores are displayed in green. The asterisk denotes a score significance of $p \leq 0.01$.

Chapter 7

Error Analysis

The compromised reliability of the human evaluation, combined with the need of more specific insights on the impact of the distinct *settings*, leads us to conduct an error analysis of the annotated responses. Constrained by limitations in time and resources, our analysis narrows down to two pivotal comparisons: a) We investigate differences between the *Combined* and *Unstructured settings* to pinpoint specific scenarios, where the addition of structured input yields superior results. For each Gricean-Maxim-inspired criterion, we analyze the instances where GODEL-COMB-HALF outperforms GODEL-UN-HALF by minimum 1 point and vise versa. b) We further scrutinize the impact of structured perspectival information by comparing GODEL-COMB-HALF against GODE-COMB-PER-HALF and GODEL-STR-SHARED against GODEL-STR-PER-SHARED. We follow the same method as in (a) with the exception that our analysis is centered on Dialogue-act and Emotion.

7.1 Godel-Comb-Half vs Godel-Un-Half

On average, we identify 4 instances per Gricean-Maxim-inspired criterion, where GODEL-COMB-HALF responses are rated higher than those produced by GODEL-UN-HALF, the analysis of which unveils the following patterns: GODEL-COMB-HALF avoids redundant repetitions of preceding dialogue content, as exemplified in Example (1)], compared to its competitor. In addition, it appears more efficient in maintaining relevance and resonating with the communicative goal in cases were input lacks essential information to address it. Example (2) represents such a scenario, where the model leverages the entity *'Tim Dolin'* for generating a response that could serve as a follow-up question, thereby compensating for the absence of knowledge to address the interlocutor's inquiry. If the model fails to utilize existing knowledge purposefully, it is inclined to relay known information, regardless of being recycled or not directly relevant to the communicative goal, rather than resorting to hallucinatory behaviors that may compromise Soundness, as Example (3) demonstrates. By contrast, in the same example, GODEL-UN-HALF falsely states that the writer and comedian *Spike Milligan* starred in *'X-Men Space Odyssey'*. Interestingly, the model is likely confused by the comic book's fictional characters *'Peter Milligan'* and *'Spyke'*, information that assumingly stems from its pretraining data, as it is not part of the OpenDialKg extract employed for finetuning.

Example:

(1) CONTEXT:

[1]Do you know Gary Chapman, tell me about some of his works

[2]Gary Chapman has worked on many things like Stories about love and feelings to composing music.

[3]Wait a second, did he write The Five Love Languages: How to Express Heartfelt Commitment to Your Mate?

KNOWLEDGE:
The Five Love Languages: How to Express Heartfelt Commitment to Your Mate written_by Gary Chapman. The Five Love Languages: How to Express Heartfelt Commitment to Your Mate written_by Gary Chapman

RESPONSE:
Godel-Comb-Half: yes he did
Godel-Un-Half: yes he did he also wrote five love languages how to express heartfelt commitment to your mate

Example:

(2) CONTEXT:

[1]I like to know more about Tess of the d'Urbervilles: A Pure Woman Faithfully Presented

[2]Sure. Tess of the d'Urbervilles: A Pure Woman Faithfully Presented is written by Tim Dolin, Thomas Hardy, and Margaret R. Higonnet. It was released in 1891.

[3]Wow sounds interesting! Can you tell me the genre please?

[4]It's a Fiction, Tragedy Novel. I'm sure you will like it.

[5]It sounds good , and I will have to check it out! Is there anything similar to this that you can suggest for me?

KNOWLEDGE:
Tess of the d'Urbervilles: A Pure Woman Faithfully Presented written_by Tim Dolin

RESPONSE:
Godel-Comb-Half: do you like tim dolin s work
Godel-Un-Half: it s speculative fiction tragedy novel it s also speculative fiction tragedy novel do you like those genres

Example:

(3) CONTEXT:

[1]I loved Monty Python and the Holy Grail. Can you recommend something similar?

[2]That was a great movie. The Life of Brian was another Monty Python movie. Do you like John Cleese and Eric Idle?

[3]I love them. I also really like Terry Gilliam and Spike Milligan. So anything that is done by them is great.

KNOWLEDGE:
Monty Python and the Holy Grail written_by Graham Chapman. Life of Brian starred_actors Graham Chapman. Life of Brian written_by Eric Idle

RESPONSE:
Godel-Comb-Half: terry gilliam is great actor spike milligan is also great actor do you like terry gilliam movies
Godel-Un-Half: terry gilliam is great actor spike milligan also starred in x men space odyssey have you seen it

Conversely, instances, where GODEL-UN-HALF outscores GODEL-COMB-HALF, amounting to 2 on average per criterion, are primarily attributed to inaccurate or highly subjective annotations, as in Example (4). Alternatively, GODEL-UN-HALF's hallucinatory tendencies create the illusion that its responses align with the content and communicative goal, as evident in Example (3) above. Consequently, it receives higher scores in criteria like Completeness and Relevance evaluated separately from Soundness.

Example:

(4) CONTEXT:

[1]I like Anne Fletcher

[2]Did you know she starred in The Proposal alongside Sandra Bullock?

KNOWLEDGE:
The Proposal starred_actors Anne Fletcher. The Proposal starred_actors Sandra Bullock

RESPONSE:
Godel-Comb-Half: no i didn t know that what else has sandra bullock been in
(Completeness score: 4.0)
Godel-Un-Half: no i didn t know that who else starred in proposal
(Completeness score: 5.0)

We consider the above patterns as the primary factors determining the positive contribution of structured input representation in Conciseness, Completeness, Relevance and Soundness. They provide preliminary evidence that the *Combined setting* augments control over the model's input processing. In essence, it guides the model in navigating dialogue content, identifying key entities and establishing associations among them to enhance informativeness, while mitigating the recycling of revisited content. On top of that, it contributes to the reliability and traceability of the generated information by creating more transparent distinctions between the available information sources namely, the dialogue context and the retraining or finetuning data, thus facilitating the tracking of how the model utilizes input.

Finally, our attempts to identify consistent model distinctions in relation to Brevity and Clarity were unsuccessful as comparable errors manifested in both models without discernible patterns or reoccurring triggers. The reasons underlying our inconclusive insights are discussed in the last Section.

7.2 Perspective vs Non-Perpective Models

Investigating the influence of perspective on the annotated model responses, we find that GODEL-COMB-PER-HALF and GODEL-STR-PER-SHARED demonstrate enhanced efficacy in conveying appropriate dialogue acts and emotions when compared to their corresponding *non-perspective models*, with the effect being particularly pronounced in relation to the *Structured setting*. The favorable impact of perspective triples is particularly evident in instances where the required response or the most recent turn conveys dialogue acts beyond declarative statements, such as commands or yes-no questions, exemplified in Example (5), and expressions of gratitude, as in Example (6). Similarly, *perspective* models yield more suitable responses to dialogue turns expressing non-neutral emotions, such as admiration (see Example (7)) and curiosity (see Example (8)). The distinguishing elements that deem the *perspective models*' responses more suitable are underlined in each example.

Example:

(5) CONTEXT:

[1]Who wrote It?

[2]The book "It" was written by Stephen King. The first time they made it into a movie, it starred Tim Curry from the Rocky Horror Picture Show. Do you like Horror movies and books?

[3]I love scary horror movies. Didn't Stephen King write The Stand as well?

[4]He did! And apparently he made an appearance in the film version of the book. he also wrote "The Shawshank Redemption" , but thats more of a prison drama.

[5]I loved Shawshank! Can you recommend more horror movies?

KNOWLEDGE:
The Stand (Horror Film) starred_actors Stephen King. The Stand (Horror Film) starred_actors Rob Lowe. user emotion love. user emotion curiosity. user dialogue-act opinion. user dialogue-act command

RESPONSE:
Godel-Comb-Half: stand is great horror film it also stars rob lowe do you like him

Godel-Comb-Per-Half: sure stephen king also starred in last stand which is horror film do you like horror movies

Example:

(6) CONTEXT:

[1]Do you know anything about the author Pearl S. Buck?

[2]Yes, he wrote The Good Earth which is a Historical period drama. Have you read it?

[3]I haven't yet. Do you know what year that was released?

[4]It was released in 1931. Another book you might like that was released the same year was The Waves.

[5]I haven

[6]I see, do they sound interesting to you? Another book I can recommend is Women in Love.

[7]Thanks! I love the historical period dramas and will definitely check it out.

KNOWLEDGE:
The Good Earth release_year 1931. The Good Earth has_genre Historical period drama. The Good Earth has_genre Historical period drama. Women in Love has_genre Historical period drama. user emotion gratitude. user emotion love. user dialogue-act opinion. user dialogue-act thanking

Godel-Str-Shared: no i haven t can you tell me more about women in love
Godel-Str-Per-Shared: you re welcome enjoy

Example:

(7) CONTEXT:

[1]Could you recommend movies similar to Kung Fu Panda?

[2]Have you seen the sequel, Kung Fu Panda 2? One of the writers, Glenn Berger, also wrote Monsters vs Aliens. Would that be something you would like?

[3]I have not seen Monsters vs Aliens. Did Keifer Sutherland lend his voice in the movie?

[4]he did! He voiced the character General W.R. Monger. He is a very talented actor. He also starred in A Few Good Men with Kevin Pollak.

KNOWLEDGE:
Monsters vs. Aliens starred_actors Kiefer Sutherland. A Few Good Men starred_actors Kiefer Sutherland. A Few Good Men starred_actors Kevin Pollak. assistant emotion neutral. assistant emotion admiration. assistant dialogue-act statement. assistant dialogue-act opinion

Godel-Str-Shared: kevin pollak starred in few good men
Godel-Str-Per-Shared: i like kevin pollak what else has he been in

Example:

(8) CONTEXT:

[1]I like books written by Charles Duhigg, can you recommend some?

[2]You may be interested in "The Power of Habit". It explores the science behind habit creation and reformation. Is that something that you would be interested in?

KNOWLEDGE:
The Power of Habit written_by Charles Duhigg. The Power of Habit in_language English. assistant emotion neutral. assistant emotion curiosity. assistant dialogue-act opinion

RESPONSE:
Godel-Comb-Half: i haven t read that one what genre is it
Godel-Comb-Per-Half: yes i d like to know more about it can you recommend any other books by charles duhigg

Focusing on the *Structured setting*, the generation of a more appropriate dialogue act or emotion, as observed in GODEL-STR-PER-SHARED responses, affects positively other Gricean-Maxim-inspired-criteria. For instance, addressing successfully interrogative prompts benefits, in turn, Completeness, Coherence, and Relevance, as showcased in Examples (9) and (10). Moreover, Brevity and Conciseness are improved for similar reasons. Example (11) demonstrates that a more appropriate dialogue-act is more straightforward in conveying meaning, while mitigating the risk of unnecessary content repetitions. Conversely, the *non-perspective model* appears less regulated and turns to unnecessary verbosity to compensate for the lack of substantive content.

Example:

(9) CONTEXT:

[1]Do you know the book, Como Agua Para Chocolate? Are you familiar with the author?

[2]Como agua para chocolate is written by Thomas Christen do you like this author?

[3]I thought the book was written by Laura Esquivel, but Thomas Christensen translated the book. I do like his books. Can you recommend one, please.

[4]It was released in 1989, the same year as The Power of one have you read it?

KNOWLEDGE:
Como agua para chocolate written_by Thomas Christensen. Como agua para

chocolate written_by Laura Esquivel. Como agua para chocolate written_by Thomas Christensen. assistant emotion curiosity. assistant dialogue-act statement

RESPONSE:
Godel-Str-Shared: i think i ll check it out thank you
Godel-Str-Per-Shared: no i haven t can you tell me more about it

Example:

(10) CONTEXT:

[1]Do you know , what the Orson Scott Card is about?

[2]Of course, Orson Scott Card wrote Xenocide, have you read that?

KNOWLEDGE:
Xenocide written_by Orson Scott Card. assistant emotion curiosity. assistant dialogue-act command

RESPONSE:
Godel-Str-Shared: what genre is xenocide
Godel-Str-Per-Shared: i haven t read xenocide what genre is it

Example:

(11) CONTEXT:

[1]Hi, have you seen American Psycho? I loved that movie and hoping you could recommend some others similar?

[2]That's a good one! Do you like it because of the genre? It's both psychological thriller and crime fiction.

[3]Yes, I love both! I really enjoyed When a Stranger Calls, so maybe something like that?

[4]I'd say check out Taxi Driver.

KNOWLEDGE:
When a Stranger Calls has_genre Psychological thriller. When a Stranger Calls has_genre Psychological thriller. Taxi Driver has_genre Psychological thriller. assistant emotion neutral. assistant dialogue-act opinion

RESPONSE:
Godel-Str-Shared: taxi driver is psychological thriller do you like it because it s thriller or because it s thriller
Godel-Str-Per-Shared: taxi driver sounds interesting who stars in it

In contrast, our efforts to discern a notable positive impact of perspective triples on Clarity yield inconclusive results, with the majority of rating differences being potentially attributed to annotator error or subjectivity, as illustrated in Example (12). While the score difference in this example may be considered insignificant for supporting this argument, the straightforward nature of the specific criterion in combination with the specificity of the annotation guidelines and the simplicity of the responses, theoretically leaves little room for subjectivity or human error. Similarly, while our analysis does not unveil solid cases, of Godel-Str-Per-Shared outperforming Godel-Str-Shared in Soundness, Example (13)—where Godel-Str-Shared model falsely places the novel in the romantic genre—supports the assumption that, when faced with limited content, the *non-perspective* model tends to primarily rely on associations between the dialogue context and its pretraining or finetuning data, thereby increasing the risk of producing inaccuracies.

Example:

(12) CONTEXT:

[1]Do you know any books written by Chuck Palahniuk?

[2]Tell-All, Survivor and Snuff

[3]What year was 'Tell-All' released?

[4]In 2010

[5]Are there any other thrillers you could recommend?

KNOWLEDGE:
2010 has_genre Thriller. X-Men 2 has_genre Thriller. user emotion curiosity. user dialogue-act yes_no_question

RESPONSE:
Godel-Str-Shared: x men 2 is thriller do you like thrillers
(Clarity score: 4.0)
Godel-Str-Per-Shared: x men 2 is thriller do you like x men movies
(Clarity score: 5.0)

Example:

(13) CONTEXT:

[1]Could you recommend some Ursula K. Le Guin books.

[2]Sure Ursula K Le Guin wrote The Tombs of Atuan, The Other Wind and The Left Hand Of Darkness.

[3]Yes I remember The Tombs of Atuan it a very interesting fiction book. When was the other two released?

[4]The Other Wind was released September 13, 2001, and The Left Hand of Darkness was in 1969.

KNOWLEDGE:
The Tombs of Atuan written_by Ursula K. Le Guin. The Tombs of Atuan written_by Ursula K. Le Guin. assistant emotion neutral. assistant dialogue-act statement

RESPONSE:
Godel-Str-Shared: tombs of atuan is romance novel written by ursula k le guin do you like romance novels
Godel-Str-Per-Shared: what genre is tombs of atuan

The observed patterns across criteria reveal a relative alignment with human judgments: we identify stronger evidence of perspective triples' positive impact on criteria, where GODEL-PER-STR-SHARED outscores more profoundly its corresponding *non-perspective model* compared to criteria, where their score difference is rather subtle.

Turning to the *Combined* setting, though human judgments rate GODEL-COMB-HALF higher than the *perspective model* in Emotion, upon a closer examination of these cases (3 in total), we find that the scoring difference is attributed to human subjectivity, as illustrated in Example (14).

Example:

(14) CONTEXT:

[1]Do you like Tom Clancy?

[2]Oh, sure, he wrote The Sum of All Fears and The Hunt for Red October.

[3]Yes The Sum of all Fears is a Thriller

KNOWLEDGE:
The Sum of All Fears written_by Tom Clancy. The Sum of All Fears has_genre Thriller. user emotion approval. user dialogue-act pos_answer

RESPONSE:
Godel-Comb-Half: yes it is thriller do you like thrillers
(Emotion score: Yes)
Godel-Comb-Per-Half: yes it is thriller do you like thrillers
(Emotion score: Part)

In contrast to the *Structured setting*, we did not discover a transparent contribution of perspective triples on Gricean-Maxim-inspired criteria for the *Combined setting*. This also aligns with human judgments, where GODEL-COMB-HALF and GODEL-COMB-PER-HALF roughly differ. We posit that this may be attributed to the inextricable role of unstructured dialogue context on response generation that likely leads to highly similar model responses, given the only input difference lies in the incorporation of perspective triples. In contrast, the additional guidance of perspective triples appears more impactful for the *Structured setting*, likely due to its lacking input.

7.3 Confounding Variables

We also identify the following deterrents for our analysis often causing inconsistencies in models' behavior and confounding our conclusions. a) As mentioned earlier, models' common underlying LLM leads to the generation of identical or closely similar responses. This is manifested more intensely in cases, where the most recent turn does not pose direct information requests, but rather has a chit-chat objective as in Example (15) below. b) The structured dialogue history often includes triples, whose content is not mentioned in the unstructured dialogue context, rather is often relevant to the reference response, due to data malformation (see Example (16)). As a result, the model employing structured information accidentally receives additional knowledge grounding, thus producing more appropriate results that match the content and communicative goal. c) The LLM's influence on responses involves a high level of randomness stemming from its architecture and its pretraining data—the latter may exhibit stronger connections to certain dialogue contents over others. d) The content of triples varies in informativeness and interaction with the unstructured input, hence models' knowledge reliance may oscillate among structured, unstructured or pretraining information for the generation of each response.

Example:

(15) CONTEXT:

[1]Can you recommend any movie with dinosaurs?

[2]Yes, have you seen Jurassic Park? It's an adventure film with dinosaurs.

[3]I have not.. who's in it?

[4]I'm not sure, but you might want to also check out The Land Before Time. It has Diana Ross in it.

KNOWLEDGE:
Eric Idle Episode segment Michael Palin. A Fish Called Wanda starred_actors Michael Palin. Time Bandits starred_actors Michael Palin. user emotion neutral. user emotion admiration. user dialogue-act open_question_factual. user dialogue-act opinion

RESPONSE:
Godel-Comb-Half: i have not seen that one who else is in it
Godel-Un-Half: i have not seen that one who else is in it

Example:

(16) CONTEXT:

[1]Do you know any books written by Chuck Palahniuk?

[2]Tell-All, Survivor and Snuff

[3]What year was 'Tell-All' released?

[4]In 2010

[5]Are there any other thrillers you could recommend?

KNOWLEDGE:
2010 has_genre Thriller. X-Men 2 has_genre Thriller

RESPONSE:
Godel-Comb-Half: x men 2 is thriller
Godel-Un-Half: do you like thrillers

Chapter 8

Additional Experiments

In light of the numerous stochastic parameters influencing the finetuned models' response generation, and, in turn, interfering with the systematic and timely comparison of our distinct representation *settings*, we undertake a series of additional experiments to strengthen our findings.

8.1 Dialogue Length Filtering

Considering that the primary focus of this study is the comparative analysis of the *qualitative settings*, any conclusions on the impact of *quantitative settings* on response generation are drawn from the automatic evaluation results alone. However, the latter might have been distorted by the substantial data skewness towards dialogues with fewer turns in their dialogue history (refer to Figure 3.1, Chapter 3). This imbalance results in a sizable input overlap among models employing distinct *quantitative settings*. For instance, a dialogue history consisting of 2 turns, is represented identically by the *Half* and *One settings* and possibly the same holds for the *All* and *Shared settings*.

To address this, we exclude from the test set any response IDs with maximum 2 turns in their dialogue history and apply the *quantitative settings* to the remaining data, ensuring distinct input configurations. We then reapply automatic metrics to the filtered test sets.

The results in Figures 8.1 and 8.2 agree with the initial model performance rankings prior to dialogue length filtering for both *non-perspective* and *perspective models*. Moreover, the score differences are amplified, confirming the reliability of our original findings despite the imbalanced data distribution.

	Model	ROUGE-L	BLEU	METEOR	BERTSCORE
L	Godel-Comb-All	**20.7791**	**5.2593**	**18.0456**	**74.9486**
T, L	Godel-Comb-Half	**22.5457**	**6.1529**	**19.8138**	**75.6106**
L	Godel-Comb-One	**21.1335**	**5.6927**	**18.3092**	**75.1387**
	Godel-Un-All	19.7081	4.7084	17.3117	74.5022
T	Godel-Un-Half	20.4888	4.8823	17.8394	74.8388
	Godel-Un-One	18.5521	4.2056	15.952	74.0748
L	Godel-Un-Shared	19.2656	4.3614	16.5588	74.2983
T	Godel-Str-All	16.4433	4.1516	14.5548	73.1187
	Godel-Str-Half	16.1671	3.8776	14.0685	73.0612
	Godel-Str-One	**15.3908**	**3.5967**	**13.3875**	**72.7253**
	Godel-Str-Shared	16.3814	3.961	14.3534	73.122

Figure 8.1: Rouge, Bleu, Meteor and BertScore performance scores of *non-perspective models* applied on the filtered test set, where each instance includes 3 or more turns in its dialogue history. For every metric the highest model scores are displayed in green from darker (1st highest score) to lighter (3rd highest score), while the lowest score is displayed in red. Models labeled with the letter "T" on their left-side perform the best across *quanTitative settings* (e.g., Godel-Comb-Half outperforms Godel-Comb-One and Godel-Comb-All). The letter "L" indicates the best performing model across *quaLitative settings* (e.g., Godel-Comb-Half outperforms Godel-Un-Half and Godel-Str-Half).

	Model	ROUGE-L	BLEU	METEOR	BERTSCORE
L	Godel-Comb-Per-All	20.7228	5.2633	18.0596	74.8555
T, L	Godel-Comb-Per-Half	**22.5824**	**6.0149**	**19.727**	**75.5857**
L	Godel-Comb-Per-One	21.2404	5.7616	18.4527	75.1886
	Godel-Str-Per-All	18.2525	**4.529**	15.7191	74.0587
	Godel-Str-Per-Half	18.1816	4.714	15.6900	74.0409
	Godel-Str-Per-One	**17.9721**	4.79	**15.2593**	73.9594
T	Godel-Str-Per-Shared	18.3513	4.683	15.8381	**73.9363**

Figure 8.2: Rouge, Bleu, Meteor and BertScore performance scores of *perspective models* applied on the filtered test set, where dialogue history includes 3 or more turns. For every metric the highest model scores are displayed in green, while the lowest score is displayed in red. Models labeled with the letter "T" on their left-side perform the best across *quanTitative settings* (e.g., Godel-Comb-Per-Half outperforms Godel-Comb-Per-One and Godel-Comb-Per-All). The letter "L" indicates the best performing model across *quaLitative settings* (e.g., Godel-Comb-Per-Half outperforms Godel-Str-Per-Half) The symbols '+', '-' and '=' preceding the numbers, denote increase, decrease or stability in performance respectively compared to the corresponding *non-perspective models*.

8.2 Adversarial Triple Imputation

To comprehensively assess the impact of structured input and the integration of perspective on response generation, we conduct the following adversarial experiment. For *non-perspective models* employing structured representations we systematically replace each triple in the test set with a random triple adhering to the syntax `[Noun, Verb, Noun]` to simulate the original triple format of `[Subject, Verb, Predicate]`. For *perspective models*, we utilize the same random triples to impute the factual ones, and an additional set for substituting the perspective ones following the formats `[Noun, 'emotion', Noun]` and `[Noun, 'dialogue-act', Noun]`. Finally, We transform the random triples into sentence-like, period-separated strings according to our original experimental setup.

Despite our initial intention to implement this experiment across the models employed in human evaluation, we replace Godel-Str-Shared and Godel-Str-Per-Shared with those utilizing half the dialogue history, as they share the highest triple overlap. This is because, the variation in the number of triples per instance caused by the *Shared setting* poses a challenge for employing the same triples in both the *Combined* and *Structured* settings, introducing an unnecessary random factor for our experiment.

Model	ROUGE-L	BLEU	METEOR	BERTSCORE
Godel-Comb-Half	22.3572 20.3999	6.4409 5.4744	20.0572 17.9501	75.7344 75.3171
Godel-Str-Half	18.0096 9.4310	4.8018 0.9327	15.9229 7.6900	74.0518 71.1446
Godel-Comb-Per-Half	-22.3446 -20.3198	-6.3344 -5.2400	-19.8251 -17.6175	-75.6293 -75.1114
Godel-Str-Per-Half	+19.2941 +10.2602	+5.304 -0.9143	+17.0117 +8.2851	+74.5835 -70.4703

Figure 8.3: Rouge, Bleu, Meteor and BertScore performance scores of *non-perspective* and *perspective models* employing the *Combined Half* and *Structured Half* representation settings and applied on the imputed structured input. The top lines represent the models' performance on the original (i.e., non-imputed input), while the bottom lines show their performance after the substitution of triples with random ones. The symbols '+', '-' and '=' preceding the scores of *perspective models* denote increase, decrease and stability in performance respectively compared to their corresponding *non-perspective models.*

Based on the findings presented in Figure 8.3, there is a noticeable decline in overall model performance following triple imputation. In*non-perspective models*, the scores of Godel-Str-Half decrease exponentially across metrics, an unsurprising outcome, considering that imputed triples constitute the sole input source for this model. By contrast, Godel-Comb-Half demonstrates a moderate drop in performance, which suggests the model's ability to circumvent low-quality input leveraging alternative sources, when available.

The same behavior is also evident in Godel-Comb-Per-Half, which experiences a marginal decline in performance. Godel-Str-Per-Half, however, shows an improvement in Meteor and Rouge scores, despite our expectation that the induction of additional noise trough the imputed perspective triples would exacerbate its performance. Upon closer examination of the instances contributing to this score difference, we conclude that the 'dialogue-act' and 'emotion' entities within the triples, present also in the model's training, direct Godel-Str-Per-Half towards more expressive and concise response templates, more likely to align with the reference, as illustrated in Examples (1) and (2). In contrast, Godel-Str-Half tends to produce declarative statements using the available triples, a pattern also observed during error analysis. The random triples are, thus, more likely to be manifested increasing the predictions' divergence from the reference.

Example:

(1) REFERENCE:
Yes! I've had my eyes on him for awhile. But Phil Davis is still my favorite.
RESPONSE:

Godel-Str-Half: Thiago Silva Sport played Mixed martial arts
(Rouge-L score: 7.40)
Godel-Str-Per-Half: i love that movie i think i ll check it out thank you for your help
(Rouge-L score: 12.12)

Example:

(2) REFERENCE:
Thats interesting. Thanks for the info. RESPONSE:
Godel-Str-Half: i don t think i ve heard of sale but i ve heard of dialogue what genre is it
(Rouge-L score: 0.0)
Godel-Str-Per-Half: that s good one i ll have to check it out thanks for your help
(Rouge-L score: 28.57)

To elucidate the reasons why this discrepancy in performance is only reflected by ROUGE-L and METEOR, we must delve into the specific attributes each metric considers. BLEU's brevity penalty renders it more resilient to the conciseness of GODEL-STR-PER-HALF responses, while its focus on precision and the presence of contingent word sequences heightens its susceptibility to triple perturbations. Similarly, BERTSCORE is more sensitive to the steep contextual changes introduced by random triples. In contrast, ROUGE-L's emphasis on recall, and METEOR's diverse heuristics for capturing semantic similarity, lead to a heightened awareness of the overlap in ancillary tokens, such as function words or vocabulary expressing perspective, between the reference and the model responses.

8.3 Knowledge Transfer

Finally, to measure the amount of dialogue history information assimilated in each model's responses, we naively count the dialogue history graph entities present in them.

In accordance with our expectations, we observe a higher graph entity preservation in the responses of models operating on structured input with the models employing the *Structured* setting surpassing those utilizing the *Combined* one. This tentatively supports the conclusions hinted by error analysis, according to which, structured input enhances the model's grounding to existing contextual information, thereby mitigating potential external disorientation arising from its pretraining or finetuning.

Models	**Entity Overlap**
Godel-Str-Per-One	0.251
Godel-Str-All	0.246
Godel-Str-Shared	0.241
Godel-Str-Half	0.239
Godel-Str-Per-Half	0.236
Godel-Str-Per-All	0.227
Godel-Str-One	0.225
Godel-Str-Per-Shared	0.221
Godel-Comb-One	0.211
Godel-Comb-Per-One	0.211
Godel-Comb-Per-Half	0.176
Godel-Comb-Half	0.175
Godel-Comb-Per-All	0.159
Godel-Comb-All	0.156
Godel-Un-All	0.151
Godel-Un-Half	0.138
Godel-Un-Shared	0.134
Godel-Un-One	0.124
Reference	0.117

Table 8.1: Knowledge overlap rates between the dialogue context and models' predicted responses. Scores range from 0 to 1 and are presented in descending order.

Chapter 9

Conclusions

This study is situated within the ever-evolving landscape of Conversational AI, motivated by the pressing need to optimize human-computer interaction, as dialogue agents increasingly permeate societal domains assuming a combination of chit-chat and task-oriented objectives. With neural architectures and, particularly, Large Language Models being their primary underlying engine, we endeavour to contribute our small part in addressing dialogue systems' recognized downsides.

Inspired by the prominent advantages of structured input in various NLG tasks, particularly showcased in the context of external knowledge grounding, we seek to explore the impact of structured dialogue representation on neural dialogue response generation. Our study is directed particularly towards LLMs, considering their leading role in the task, yet their acknowledged limitations including the generation of hallucinatory information, their elusive preprocessing and learning behaviors, and the lack of a standardized and insightful evaluation framework. We argue that a structured representation of dialogue can contribute to the reliability, controllability and interpretability of such systems in generating responses, while also elevating the overall response quality. Our analysis centers on open-domain goal-directed dialogue, recognizing its dual character as a fundamental attribute for crafting engaging and utilitarian modern dialogue agents.

We empirically explore our hypothesis by introducing a total of 3 *qualitative* and 4 *quantitative settings* for representing dialogue history. The former pertains to the type of dialogue representation and is distinguished into *Structured, Unstructured* and *Combined*. The latter determines the number of turns considered in dialogue history, classified as *All, Half, One* and *Shared*. We combine these *settings* to formulate 11 distinct input configurations of the OpenDialKg dataset (Moon et al., 2019), which we use to finetune 11 *non-perspective models* leveraging GODEL (Peng et al., 2022), a transformer-based LLM. Additionally, arguing that the factual triples constituting the structured representation of dialogues are inadequate for capturing their intricate properties, we enhance structured representations with perspectival information pertaining to dialogue-acts and emotions and use it to finetune 7 additional *perspective models*. To assess response quality we employ four standardized automatic metrics, namely ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTSCORE (Zhang et al., 2019), complemented by a novel manual evaluation framework. The latter utilizes 7 finegrained criteria inspired by the Gricean Maxims (Grice, 1975), and 3 exploratory ones assessing the expression of Dialogue-act and Emotion, as well as the fulfillment of the Communicative goal. Finally, to deepen our understanding and fortify the reliability of our initial findings, we incorporate three supplementary experiments. These include triple perturbations, the tracing of knowledge transfer from dialogue histories to model responses, and the filtering of dialogue history quantity to distinguish more effectively the impact of *quantitiative settings*.

In essence, our study does not seek to attain state-of-the art performance in dialogue response generation. Rather, its core objective is exploratory, summarized in the following research question and subquestions:

How does a structured representation of dialogue history impact the quality of neural response generation, as measured by standardized automatic NLG metrics and a manual evaluation inspired by the Gricean Maxims?

- How does the response quality differ, when dialogue history representation is driven by each of the introduced *qualitative settings*?
- How do the introduced *quantitative settings* interact with the aforementioned qualitative ones in generating dialogue responses as reflected by response quality?
- How does the incorporation of additional perspectival information into the structured representation of dialogue history affects response quality?

Overall, our automatic and manual evaluation along with our additional experiments reveal that integrating structured dialogue history representation into unstructured dialogue context has an amplifying effect on model performance. Yet, when employed in isolation, structured representation yields inferior results, as expected, given its inability to comprehensively capture the intricate spectrum of conversational knowledge, such as diverse perspectival information, essential for fostering natural, clear and engaging communication. Notably, the incorporation of perspective triples occupies a reinforcing role for the models utilizing solely structured input, as displayed across all facets of our study. Conversely, its impact on models operating on both input types is deemed negligible by automatic metrics, while manual evaluation and error analysis deem it more positive for certain aspects of response quality.

Regarding history length, automatic metrics unveil that considering half of the dialogue history turns serves a solid initial benchmark across all representation types. However, models integrating both structured and unstructured input appear less 'information hungry' performing optimally on smaller history quantities, compared to models relying solely on one input type. The latter necessitate longer input for context extraction, thereby running a higher risk of incorporating irrelevant information. Furthermore judging from the *Shared setting* scores, filtering history to include only turns pertinent to the current prompt proves advantageous for model performance.

Our introduced manual evaluation poses considerable challenges. Initial inter-annotator agreement appears modest, though largely improves after addressing noisy ratings with Soundness, Coherence and Completeness achieving an acceptable Krippendorff's α exceeding 0.6. A detailed analysis of disagreement exposes an alarming level of human error and extensive space for subjective interpretations, compromising the reliability of our human evaluation results. Importantly, problematic instances are isolated to specific scenarios, providing prospects for future refinement. In addition, we manage to overcome the interpretability challenges of Krippendorff's α arising mainly from the noisy ratings, by consulting it in combination with the Mean Absolute Error calculated between score pairs. This way, we are led to more systematic insights into the nature of disagreement and the scores it concerns, that could be otherwise attained through a timely manual inspection.

Ultimately, human ratings exhibit an overall weak correlation with automatic metrics with peak values oscillating around 0.3 for Completeness, Relevance, and Coherence, while Dialogue-act and Communicative goal display an inverse correlation. Significance scores underscore the need for a larger sample size in human evaluations to achieve robust correlation judgments for most criteria. However, despite the modest response-level correlations automatic and manual techniques broadly concur in model ranking, with the exception that human judgments deem the incorporation of perspective slightly more fruitful for the *Combined setting*. These observations highlight that automatic metrics hinge on distinct response properties both compared to each other and to human evaluation criteria. Indeed, during our additional experiments, it becomes evident that BLEU and BERTSCORE assign greater importance to core content of responses, while ROUGE and METEOR adopt a more holistic approach, considering peripheral details as well.

We shed light on the previously obscure findings through a thorough, yet not exhaustive, error analysis. Specifically, a meticulous review of the models' responses unveils that instances where the *Unstructured setting* outperforms the *Combined* one are predominantly attributed to inaccurate annotations. Moreover, it resolves the debate surrounding the contribution of perspective in the *Combined setting* by offering evidence of its positive influence on the expression of Emotion and Dialogue-acts. In summary, cumulative evidence from the examined responses, the automatic and human evaluation and the supplementary experiments lay a modest foundation for the following preliminary conclusions regarding the impact of structured dialogue representation on response generation.

- Integrating a structured representation into raw dialogue context augments the **controllability** over models' content processing directing attention towards core content over ancillary conversational details. This claim is substantiated by error analysis and the relative automatic score adjacency between the models operating on exclusively structured versus unstructured inputs, especially after the addition of perspective.

- Augmenting unstructured dialogue input with its structured counterpart contributes to effective context modeling, thereby **satisfying** models' **information requirements**. This is evidenced on two fronts: Firstly, models operating on both input types are more efficient in handling smaller amounts of history. Secondly, while perspective triples add to the deprecated informativeness of models operating solely on structured input, they have an adversely effect on those utilizing combined input, likely because the model's information needs are already addressed.

- Improving the informativeness of the input (see previous point) enhances models' **resourcefulness**, as observed in their reliance on triple entities for generating appropriate responses in low-information settings. This resourcefulness is further demonstrated during the imputation of triples with random ones, with models relying on both input types showing resilience by resorting to the unperturbed unstructured context.

- Structured input mitigates models' hallucinatory behavior by intensifying grounding in existing contextual information, as revealed in error analysis, thereby enhancing **reliability**. On top of that, it assists in **tracing the origins of the information** included in responses, establishing connections between the model's pretraining and finetuning data or the dialogue context.

- Enhancing the structured representation of dialogue with perspectival input improves models' ability to express and respond appropriately to dialogue-acts and emotions, contributing to response **naturalness**, particularly in cases involving non-declarative dialogue acts and non-neutral emotions.

- Structured perspectival information has a holistic impact on response quality, especially pronounced when dialogue is solely represented structurally and informativeness is limited.

- The above strengths are interrelated contributing collectively to the **overall quality** of the generated responses including relevance, soundness, coherence, conciseness and completeness, by establishing an enhanced awareness of the dialogue content.

We find the identification of more robust clear cut-distinctions among our introduced *settings* rather challenging, due to a multitude of confounding factors including the simplistic nature of the dialogues, the limited scope of the available structured representations, the inaccuracies in the mapping between dialogue turns and triples, the high error rate of the human evaluation process, as well as the inherent randomness and pretraining data biases stemming from the use of an LLM. This plethora of variables affecting model performance beyond our introduced *qualitative* and *quantitative settings* introduces a considerable degree of randomness and a lack of control into our study that is not possible to systematically investigate. It also implies that achieving greater reliability and insightfulness in our findings requires a meticulous examination of a larger number of instances. Nevertheless, our exploratory study successfully addresses our research questions, extracting preliminary observations that support the positive impact of structured dialogue representation and structured perspective knowledge, which can be considered an encouraging starting point for future analysis.

On that account potential future endeavors include the reimplementation of the manual evaluation process on a larger scale following a refinement of the annotation guidelines and the conduction of a trial annotation process. Other directions involve a comparative analysis of how various Large Language Models (LLMs) interact with our introduced representation *settings*, given their distinct underlying architectures and pretraining. Most importantly, we consider the creation of a novel dataset tailored specifically for this task to be an invaluable contribution towards the comprehensive exploration of the impact of structured dialogue representation across various setups, such as through the exploitation of graph embeddings integrated on a simpler neural encoder-decoder framework.

Appendix A

Perspective Extraction

A.1 Emotion Classification

Positive		Negative		Ambiguous
admiration 👏	joy 😃	anger 😡	grief 😢	confusion 😕
amusement 😂	love ❤️	annoyance 😒	nervousness 😬	curiosity 🤔
approval 👍	optimism 🤞	disappointment	remorse 😔	realization 💡
caring 🤗	pride 😌	disapproval 👎	sadness 😞	surprise 😲
desire 😍	relief 😅	disgust 🤮		
excitement 🤩		embarrassment 😳		
gratitude 🙏		fear 😨		

Figure A.1: The emotion labels (Alon and Ko, 2021) used by the emotion classifier applied on every dialogue turn

A.2 Dialogue-Act Classification

Dialog Act - Semantic request			
Dialog Act Tag	**Description**	**Example**	**Count in user utterances (single label only)**
factual question	factual questions	How old is Tom Cruise; How's the weather today	360
opinion question	opinionated questions	What's your favorite book; what do you think of disney movies	236
yes/no question	yes or no questions	Do you like pizza; did you watch the game last night	325
task command	commands/requests (can be in a question format) for some actions that may be different from the ongoing conversation	can i ask you a question; let's talk about the immigration policy; repeat	651
invalid command	general device/system commands that cannot be handled by the social bot	show me a picture; cook food for me	87
appreciation	appreciation towards the previous utterance	that's cool; that's really awesome	201
general opinion	personal view with polarized sentiment	dogs are adorable; (A: How do you like Tom) B: i think he is great	2157
complaint	complaint about the response from another party	I can't hear you; what are you talking about; you didn't answer my question	239
comment	comments on the response from another conversation party	(A: my friend thinks we live in the matrix) B1: she is probably right; B2: you are joking, right; B3: i agree; (A: ... we can learn a lot from movies ...) B: there is a lot to learn; (A: He is the best dancer after michael jackson. What do you think) B: michael jackson	430
statement non-opinion	factual information	I have a dog named Max; I am 10 years old; (A: what movie have you seen recently) B: the avengers	1717
other answer	answers that are neither positive or negative	I don't know; i don't have a favorite; (A: do you like listening to music) B: occasionally	428
positive answer	positive_answers	yes; sure; i think so; why not	1278
negative answer	negative response to a previous question	no; not really; nothing right now	867

Figure A.2: The dialogue-act labels representing semantic requests and their metadata, which are employed by the MIDAS dialogue act classifier (Yu and Yu, 2019) applied on every dialogue turn

Dialog Act - Functional request			
Dialog Act Tag	**Description**	**Example**	**Count in user utterances (single label only)**
abandon	not a complete sentence	So uh; I think; can we	440
nonsense	utterances that do not make sense to humans	he all out	129
hold	a pause before saying something	let me see; well	272
opening	opening of a conversation	hello my name is tom; hi;	
closing	closing of a conversation	nice talking to you; goodbye	540
thanks	expression of thankfulness	thank you	80
back-channeling	acknowledgement to the previous utterance	Uh-huh; (A: i learned that ...) B: okay/yeah/right/really?	427
apology	apology	I'm sorry	29
apology response	response to apologies	That's all right	6
other	utterances that cannot be assigned to other tags		12

Figure A.3: The dialogue-act labels representing functional requests and their metadata, which are employed by the MIDAS dialogue act classifier (Yu and Yu, 2019) applied on every dialogue turn

Appendix B

Annotation Guidelines

Annotation Guidelines

Evaluate each dialogue response given its dialogue context (i.e., dialogue history) and the following criteria, distinguished in two categories.

Please, consider that the responses have been tokenized, lowercased and that punctuation and articles have been removed, therefore, these are not mistakes made by the models.

CATEGORY I

- the properties are evaluated on a Likert scale *(1-5)*.
- if score equals *0*, put *1* instead
- if score has first decimal < *.5* round down
- if score has first decimal >= *.5* round up

Soundness

The response should contain conceptually logical information that is likely to be true based on common sense and factual knowledge. If you are not certain about easy-to-retrieve information, as in (b), please search online. You don't have to search for information that is challenging to find, as in (c).

5 all statements are true

4 the truthfulness of one statement is obscure

3 one statement is not true, but there are other true statements

2 two or more statements are not true, but there are other true statements

1 none of the statements are true

eg.

a) RESPONSE: "*Michael Jackson is a country.*" ***(1)***

b) RESPONSE: *"Michael Jackson is a performer and a politician."* ***(3)***

c) RESPONSE: *"Michael Jackson went to France 30 years ago."* ***(4)***

d) RESPONSE: "*I like Michael Jackson!"* ***(5)***

Conciseness

The response should not provide more content than necessary for the communicative goal to be addressed and the meaning to be conveyed.

NOTE: The communicative goal does not have to be achieved (see d).

5 no redundant statements

4 one redundant statement

3 two redundant statements

2 three redundant statements

1 four or more redundant statements

eg.

HISTORY: *"When did Michael Jackson die?"*

a) RESPONSE: *"Michael Jackson died in 2009."* ***(5)***

b) RESPONSE: *"Michael Jackson died from intoxication."* ***(4)***

c) RESPONSE: *"Michael Jackson died from intoxication in 2009."* ***(4)***

d) RESPONSE: *"I don't know"* ***(5)***

Completeness

The response should provide all the information necessary for the communicative goal to be addressed and its meaning to be conveyed.

.

NOTE: A complete response does not guarantee that the communicative goal is achieved. In (c) the response is complete and scores 5, even though the communicative goal (i.e., knowledge acquisition) is not reached.

Completeness = (the amount of <u>necessary</u> info stated / the amount of info we judge is necessary) * 5

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: *"Yes, I could totally recommend it."* ***(3)*** ½ * 5 = 2.5 = 3
(It doesn't mention who is starring)

b) RESPONSE: *"I don't know."* ***(1)*** 0/2 * 5 = 0 = 1
(It doesn't mention what the speaker doesn't know)

c) RESPONSE *"I would totally recommend it. I don't know who is starring."* ***(5)*** 2/2 * 5 = 5

Relevance

The response should relate to the conversation history and the communicative goal, as you perceive it.

NOTE: The communicative goal does not have to be achieved, but it does need to be considered for the generation of the response.

5 relevant to most recent turn and communicative goal of the entire history with specific details

4 relevant only to the most recent turn with moderate specificity and likely containing a rather generic cue *(eg. Do you like X?)*

3 a very generic response, but still applicable *(eg. I don't know).*

2 only thematically (i.e., topic-based) relevant to the most recent turn, but not the communicative goal

1 not relevant at all

eg.

HISTORY:

- *"Do you know Jordan Smith?"*
- *"I think I saw him on TV. He's a golfer born in Dallas"*
- *"Do you know any other athletes?"*
- *"What about Rohit Sharma? He plays for the Mumbai Indians."*

a) RESPONSE: *"He's a great player! He also plays in the national team."* ***(5)***

(Relevant to most recent turn. Relevant to the communicative goal of the entire history i.e., sports chit chat and knowledge exchange. The second sentence adds specificity).

b) RESPONSE: *"I haven't heard of him. Do you like the Mumbai Indians?"* ***(4)***

c) RESPONSE: "Rohit Sharma is an athlete." ***(2)***

Clarity

The response should NOT be:

- semantically ambiguous (its meaning allows more than one interpretation)
- syntactically ambiguous (its syntax allows more than one interpretation)
- semantically obscure (the concepts or words do not convey a clear meaning, sound likely unnatural and are hard to understand)
- syntactically obscure (the structure and grammar are complex and/or unnatural and require careful parsing to interpret the response.)

5 zero undesired properties present

4 one undesired property present

3 two undesired properties present

2 three undesired properties present

1 all undesired properties present

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: "*I don't know!"* ***(4)***

(Semantically ambiguous)

b) RESPONSE: "*Recommend the stars!"* ***(3)***

(Syntactically ambiguous, semantically obscure)

c) RESPONSE: "A performance to like." ***(3)***

(Semantically ambiguous, syntactically obscure)

Brevity

The response should not contain any unnecessary verbalizations, such as word or phrase repetitions. The response should demonstrate ability to use anaphoric expressions either within itself or in relation to the previous context

NOTE: Brevity should be distinguished from Conciseness. Conciseness refers to the conceptual content of the response, while Brevity to the lexical content of the response.

5 no unnecessary verbosity

4 one unnecessary, but grammatical verbosity

3 two unnecessary, but grammatical, verbosities

2 ungrammatical unnecessary verbosities, but the meaning can still be discerned

1 random ungrammatical and unnecessary verbosities that hinder interpretation

eg.

HISTORY: *"Would you recommend this movie? Who is starring there?"*

a) RESPONSE: *"I don't know who is starring there in this movie."* ***(3)***

(Two grammatical verbosities: "there" and "in this movie").

b) RESPONSE: *"I would recommend this movie and this movie."* ***(2)***

(The second "this movie" is an ungrammatical verbosity, given that there is no context suggesting the existence of a second movie)

c) RESPONSE: *"I would recommend this movie and on".* ***(1)***

("on" is a random, ungrammatical verbosity that hinders interpretation).

Coherence

The information presented in the response should be semantically and syntactically connected to each other and the most recent history turn in a logical order.

5 strong coherence both within the response and in relation to the previous turn

4 weak coherence either within the response OR in relation to the previous turn

3 coherence is lacking within the response OR in relation to the previous turn, but there is not a confusion in interpretation

2 coherence is lacking within the response OR in relation to the previous turn, and difficult to interpret the meaning

1 no coherence

eg.

HISTORY: *"Do you know Selena Gomez?"*

a) RESPONSE: *"Yes, she's an American singer. Do you like rock music?"* ***(4)***
(The second sentence displays weak semantic coherence in relation to the first sentence since Selena Gomez belongs in the pop genre).

b) RESPONSE: "*I've never heard of Katy Perry! What genre of music does she sing?"* ***(2)***
(The first sentence lacks coherence in relation to the last history turn, causing confusion in interpretation).

c) RESPONSE: "*Tell me a song of hers! I like "Liar."* ***(3)***
(The logical connection between the two sentences and between the response and the last history turn is lacking but the meaning can still be conveyed. The first sentence in the response suggests that the speaker is not sure if they know the singer and ask for details. However, the second sentence suggests they already know the singer).

CATEGORY II

- The properties are evaluated using the following fixed categorical values:

Y (yes)
N (no)
P (part)

Perspective: Dialogue Act

Please find the dialogue act classes used in this work on Appendix A.2

NOTE: -Always take into account only the given context, and not other factors that might influence the dialogue act in the real world, such as previous conversations or the physical context. For instance, the dialogue act in (c) might fit in a real-world setting (eg. if the speaker has been asked the same question repeatedly), but it does not match the given dialogue context.
-The most neutral dialogue act type is <u>*statement-non opinion*</u> and is likely to fit in most contexts, but without being the perfect candidate (see e). It is up to your judgment to decide whether this type is appropriate, sufficient and natural given the context.

Y (yes) the response displays an appropriate dialogue act, given the dialogue history
N (no) the response displays an inappropriate dialogue act, given the dialogue history
P(part) only part of the response displays an appropriate dialogue act
eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: *"I am so sorry!"* ***(N)***
(Apology)

b) RESPONSE: *"You're going to love it. Leo is starring."* ***(Y)***

(General Opinion + Statement-non opinion)

c) RESPONSE: *"Not again! Really?"* ***(N)***

(Complaint)

d) RESPONSE: *"This movie is directed by Martin Scorsese starring Leonardo DiCaprio."* ***(P)***

(The *statement-non-opinion* dialogue act addressing the second sentence of the history turn, is appropriate. There is no dialogue act directed to the first sentence of the history turn).

e) RESPONSE: *"This movie has received good reviews."* ***(N)***

(The *statement-non-opinion* dialogue act addressing the first sentence of the history turn is not appropriate. There is no dialogue act directed to the first sentence of the history turn.)

Perspective: Emotion

Please, find the emotion classes used in this work on Appendix A.1

NOTE: - The most neutral emotion type is <u>*neutral*</u> and is likely to fit in most contexts, but without being the perfect candidate. It is up to your judgment to decide whether this type is appropriate, sufficient and natural given the context.
- Always take into account only the given context, and not other factors that might influence the emotion in the real world, such as previous conversations or the physical context.

Y (yes) the response displays a relevant emotion

N (no) the response displays an irrelevant emotion

P(part) only part of the response displays a relevant emotion

eg.

HISTORY: *"Would you recommend this movie? Who is starring?"*

a) RESPONSE: *"I didn't really like it."* ***(P)***

(The emotion addressing the first sentence of the dialogue history is appropriate, but the second sentence is not addressed emotionally).

Communicative Goal

Knowledge exchange, knowledge acquisition and chit-chat on a specific topic are the most frequent communicative goals in the data.

NOTE: - For a response to be labeled with Y, the response needs to achieve the communicative goal, not just be relevant to it.

Y (yes) the communicative goal is achieved

N (no) the communicative goal is not achieved

P(part) the communicative goal is partially achieved or not all communicative goals are achieved.

eg.

	HISTORY:	*"Would you recommend this movie? Who is starring?"*	
a)	RESPONSE:	"*Johnny Depp is starring in the Pirates of the Caribbean. I like him."* (The communicative goal expressed by the first sentence in the dialogue history is not achieved).	***(P)***
b)	RESPONSE:	*"I wouldn't recommend. I don't remember the actor's name."* (The second part of the response is relevant to the second question of the dialogue history, but the goal i.e., acquisition of knowledge is not achieved).	***(P)***

Appendix C

Human Evaluation

C.1 Outlier Investigation

Below we provide an error analysis of the most frequent outlier pairs per category in our annotation set, displayed in Figure C.1. For each pair we attempt to identify the most divergent score and the potential reasons underlying disagreement.

Criteria	# Outliers	1st Most Common Outlier Pair	2nd Most Common Outlier Pair	3rd Most Common Outlier Pair
Soundness	29	(1-4) x 12	(1-5) x 6	(5-2) x 5
Conciseness	11	(1-4) x 4	(2-5) x 4	(5-2) x 2
Completeness	25	(1-5) x 16	(1-4) x 4	(5-2) x 3
Relevance	21	(1-4) x 7	(5-2) x 5	(2-5) x 5
Clarity	21	(5-1) x 7	(2-5) x 6	(1-5) x 4
Brevity	9	(5-2) x 4	(2-5) x 3	(1-5) x 1
Coherence	27	(5-2) x 13	(1-5) x 5	(2-5) x 5

Figure C.1: The amount of outliers and 3 most frequent outlier pairs per human evaluation criterion.

C.1.1 Soundness

Please, note that the validity of the statements in the responses is evaluated through an internet search. If relevant information cannot be easily traced, the statement is considered obscured.

PAIR: d-e — (1-4)

[1] I loved Monty Python and the Holy Grail. Can you recommend something similar?

[2] That was a great movie. The Life of Brian was another Monty Python movie. Do you like John Cleese and Eric Idle?

[3] I love them. I also really like Terry Gilliam and Spike Milligan. So anything that is done by them is great.

[4] A Fish Called Wanda and Time Bandits are not Python movies but along the same lines.

[5] Time Bandits sounds good. Who's in that?

RESPONSE: i don t know much about time bandits but i do know that it was written by tom hanks do you like tom hanks movies

DUBIOUS SCORE: **4**

ANNOTATOR: **e**

EXPLANATION: None of the statements are true

PAIR: b-c — (1-4)

[1] Do you know who wrote I know why the caged bird sings?

[2] Sure! It was written by Maya Angelou. It is an autobiographical novel. Are you a fan of those?

[3] I am definitely a fan of Maya Angelou, can you recommend more by her?

RESPONSE: sure she also wrote secret life of man it is autobiographical novel are you fan of autobiographical novels

DUBIOUS SCORE: **4**

ANNOTATOR: **c**

EXPLANATION: None of the statements are true

PAIR: b-c — (1-4)

[1] Do you know who wrote I know why the caged bird sings?

[2] Sure! It was written by Maya Angelou. It is an autobiographical novel. Are you a fan of those?

[3] I am definitely a fan of Maya Angelou, can you recommend more by her?

RESPONSE: sure she also wrote great gatsby it is autobiographical novel are you fan of autobiographies

DUBIOUS SCORE: **4**

ANNOTATOR: **c**

EXPLANATION: None of the statements are true

PAIR: b-c — (1-4)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: runaway bunny was written by clement hurd

DUBIOUS SCORE: **Subjective**

ANNOTATOR: -

EXPLANATION: Annotator (b) interpreted it as completely wrong, while Annotator (c) interpreted it as obscure, since Clement Hurd was the illustrator of the book.

PAIR: b-c — (1-4)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: clement hurd

DUBIOUS SCORE: **Subjective**

ANNOTATOR: -

EXPLANATION: Annotator (b) interpreted it as completely wrong, while Annotator (c) interpreted it as obscure, since Clement Hurd was the illustrator of the book.

PAIR: b-c — (1-4)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: clement hurd wrote runaway bunny

DUBIOUS SCORE: **Subjective**

ANNOTATOR: -

EXPLANATION: Annotator (b) interpreted it as completely wrong, while Annotator (c) interpreted it as obscure, since Clement Hurd was the illustrator of the book.

PAIR: b-c — (1-4)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: runaway bunny was written by charles dickens

DUBIOUS SCORE: **4**

ANNOTATOR: **c**

EXPLANATION: None of the statements are true

PAIR: b-c — (1-4)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: clement hurd wrote runaway bunny

DUBIOUS SCORE: **Subjective**

ANNOTATOR: -

EXPLANATION: Annotator (b) interpreted it as completely wrong, while Annotator (c) interpreted it as obscure, since Clement Hurd was the illustrator of the book.

PAIR: b-c — (1-4)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: it was written by clement hurd and was released in 1942

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: One statement is not true but there are other true statements. The correct score would be 3, therefore score 1 is the least suitable.

PAIR: a-d — (1-4)

[1] Could you recommend something like the book Reflected in You?

[2] Do you enjoy the writing of this book's author Sylvia Day?

[3] Yes! What else has she written?

[4] Some of Sylvia Day's books include Seven Years to Sin and Reflected in You. Have you heard of any of these?

[5] What genres do some of he books fall under?

RESPONSE: afterburn is drama do you like that genre

DUBIOUS SCORE: **1**

ANNOTATOR: **a**:

EXPLANATION: All statements are correct. The correct score would be 5, therefore 1 is the least suitable.

PAIR: a-d — (1-4)

[1] Could you recommend something like the book Reflected in You?

[2] Do you enjoy the writing of this book's author Sylvia Day?

[3] Yes! What else has she written?

[4] Some of Sylvia Day's books include Seven Years to Sin and Reflected in You. Have you heard of any of these?

[5] What genres do some of he books fall under?

RESPONSE: afterburn is drama do you like dramas

DUBIOUS SCORE: **1**

ANNOTATOR: **a**:

EXPLANATION: All statements are correct. The correct score would be 5, therefore 1 is the least suitable.

PAIR: a-d — (1-4)

[1] Do you know anything about the author Pearl S. Buck?

[2] Yes, he wrote The Good Earth which is a Historical period drama. Have you read it?

[3] I haven't yet. Do you know what year that was released?

[4] It was released in 1931. Another book you might like that was released the same year was The Waves.

[5] I haven

[6] I see, do they sound interesting to you? Another book I can recommend is Women in Love.

[7] Thanks! I love the historical period dramas and will definitely check it out.

RESPONSE: no i haven t can you tell me more about women in love

DUBIOUS SCORE: **Subjective**

ANNOTATOR: -

EXPLANATION: 'no i haven t' is the only statement. Annotator (a) evaluated it as obscure, while annotator (d) considered it completely false. A refinement of the guidelines for Soundness to account for incoherent responses could potentially mitigate such disagreement.

C.1.2 Conciseness

PAIR: a-b — (1-4)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: sure uglies trilogy uglies 1 3 was released in 2007 it is interesting collection definitely worth checking out

DUBIOUS SCORE: **Subjectve**

ANNOTATOR: -

EXPLANATION: In terms of factual knowledge 'uglies trilogy' the only information necessary. Annotator (a) considered the remaining information as redundant, while annotator (b) as necessary for addressing the communicative goal of chit chat.

PAIR: a-b — (1-4)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: sure uglies trilogy uglies 1 3 was released in 2007 it is interesting collection worth checking out

DUBIOUS SCORE: **Subjectve**

ANNOTATOR: -

EXPLANATION: In terms of factual knowledge 'uglies trilogy' the only information necessary. Annotator (a) considered the remaining information as redundant, while annotator (b) as necessary for addressing the communicative goal of chit chat.

PAIR: a-b — (1-4)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: sure uglies trilogy uglies 1 3 was released in 2007 it is interesting collection definitely worth checking out

DUBIOUS SCORE: **Subjectve**

ANNOTATOR: -

EXPLANATION: In terms of factual knowledge 'uglies trilogy' the only information necessary. Annotator (a) considered the remaining information as redundant, while annotator (b) as necessary for addressing the communicative goal of chit chat.

PAIR: a-b — (1-4)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: yes it was released in 2007

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: The response includes one redundant statement, therefore score 1 is the least suitable.

C.1.3 Completeness

PAIR: a-b — (1-5)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: sure uglies trilogy uglies 1 3 was released in 2007 it is interesting collection definitely worth checking out

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Since the response addresses the main question, while also offering more context for facilitating chitchat, it seems counterintuitive scoring it with 1.

PAIR: a-b — (1-5)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: sure uglies trilogy uglies 1 3 was released in 2007 it is interesting collection worth checking out

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Since the response addresses the main question, while also offering more context for facilitating chitchat, it seems counterintuitive scoring it with 1.

PAIR: a-b — (1-5)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, #1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: sure uglies trilogy uglies 1 3 was released in 2007 it is interesting collection definitely worth checking out

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Since the response addresses the main question, while also offering more context for facilitating chitchat, it seems counterintuitive scoring it with 1.

PAIR: b-c — (1-5)

[1] What do you think about Davis Levithan?

[2] David Levithan wrote Two Boys Kissing, The Realm of Possibility, Nick Norah's Infinite Playlist among others. Do you like his work?

[3] I have seen the movie Nick and Norah's Infinite Playlist. It was pretty good. Have you seen it?

[4] Yeah, I thought it was funny. Do you like comedies?

[5] I do! I really liked What to Expect When You're Expecting. I was pregnant when it came out and thought it was pretty relatable.

RESPONSE: what to expect when you re expecting is also comedy

DUBIOUS SCORE: **5**

ANNOTATOR: **c**

EXPLANATION: The information does not provide any information necessary to address the communicative goal. Its content is already mentioned in the dialogue history. Therefore, score 5 is the least appropriate.

PAIR: b-c — (1-5)

[1] Do you have movies that are kinda like Divergent?

[2] Insurgent.

RESPONSE: who else is in that movie

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: The response provides an opportunity to continue the chit-chat aligning with the communicative goal of the dialogue. Therefore it can be considered necessary rendering score 1 the least suitable. However, such instances rely heavily on the annotator's perception of the communicative goal.

PAIR: b-c — (1-5)

[1] Do you have movies that are kinda like Divergent?

[2] Insurgent.

RESPONSE: who else starred in insurgent

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: The response provides an opportunity to continue the chit-chat aligning with the communicative goal of the dialogue. Therefore it can be considered necessary rendering score 1 the least suitable. However, such instances rely heavily on the annotator's perception of the communicative goal.

PAIR: c-a — (1-5)

[1] Could you recommend some similar movies to Public Enemies?

[2] Sure! Do you like the writers Ronan Bennett or Michael Mann at all?

[3] I really liked the actor Christian Bale and Johnny Depp in this movie but the story line was great too.

RESPONSE: yes i love johnny depp what else has he been in

DUBIOUS SCORE: **1**

ANNOTATOR: **c**

EXPLANATION: The response provides an opportunity to continue the chit-chat aligning with the communicative goal of the dialogue. Therefore it can be considered necessary rendering score 1 the least suitable.

PAIR: d-e — (1-5)

[1] I loved Monty Python and the Holy Grail. Can you recommend something similar?

[2] That was a great movie. The Life of Brian was another Monty Python movie. Do you like John Cleese and Eric Idle?

[3] I love them. I also really like Terry Gilliam and Spike Milligan. So anything that is done by them is great.

[4] A Fish Called Wanda and Time Bandits are not Python movies but along the same lines.

[5] Time Bandits sounds good. Who's in that?

RESPONSE: i don t know much about time bandits but i do know that it was written by tom hanks do you like tom hanks movies

DUBIOUS SCORE: **1**

ANNOTATOR: **d**

EXPLANATION: The first statement addresses the most recent question in dialogue history, while the remaining statements can be considered useful for facilitating chit-chat rendering score 1 the least suitable.

PAIR: d-e — (1-5)

[1] Could you recommend anything by Joseph Gordon-Levitt?

[2] Sure. Joseph Gordon-Levitt starred in Looper alongside Bruce Willis. Have you seen that one?

[3] No I haven't, Can you tell me something about Looper?

[4] Looper is a Sci-Fi film directed and written by Rian Johnson. It also stars Emily Blunt.

[5] What genre is it?

[6] It's a Sci-Fi/Action/Thriller. Do you like any of those genres?

[7] What movies would you recommend that's in the thriller genre?

[8] You might like Shark Night, starring Katharine McPhee, or Zodiac.

[9] Who else starred in Shark Night?

RESPONSE: i m not sure who else starred in shark night

DUBIOUS SCORE: **1**

ANNOTATOR: **d**

EXPLANATION: The first statement addresses the most recent question and thus the communicative goal without necessarily achieving it. Therefore, score 1 is considered the least suitable.

PAIR: b-c — (1-5)

[1] I love harry potter and the order of the phoenix. Can you tell me more movies like this

[2] Definitely. It's part of a series. Movies like it are Harry Potter and the Goblet of Fire.

RESPONSE: i haven t what else has she written

DUBIOUS SCORE: **5**

ANNOTATOR: **c**

EXPLANATION: The information included in the response is not necessary neither for conveying its meaning or addressing the communicative goal, rendering score 5 the least suitable.

PAIR: b-c — (1-5)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: runaway bunny was written by clement hurd

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: Though the information provided is inaccurate, the response addresses successfully the most recent question, thus rendering score 1 the least suitable.

PAIR: b-c — (1-5)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: clement hurd

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: Though the information provided is inaccurate, the response addresses successfully the most recent question, thus rendering score 1 the least suitable.

PAIR: b-c — (1-5)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: clement hurd wrote runaway bunny

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: Though the information provided is inaccurate, the response addresses successfully the most recent question, thus rendering score 1 the least suitable.

PAIR: b-c — (1-5)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: runaway bunny was written by charles dickens

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: Though the information provided is inaccurate, the response addresses successfully the most recent question, thus rendering score 1 the least suitable.

PAIR: b-c — (1-5)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: clement hurd wrote runaway bunny

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: Though the information provided is inaccurate, the response addresses successfully the most recent question, thus rendering score 1 the least suitable.

PAIR: b-c — (1-5)

[1] Could you recommend something written by Matthew Ward?

[2] Sure. He wrote L'Etranger

[3] I see. When it was released? Is it from the last century?

[4] It was released in 1942

[5] Can you tell me the genre as well?

[6] The genre is childrens literature. Another great one released that year was Runaway Bunny

[7] Great! Seems like a good year for the genre. Who wrote it?

RESPONSE: it was written by clement hurd and was released in 1942

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: Though the information provided is inaccurate, the response addresses successfully the most recent question, thus rendering score 1 the least suitable.

C.1.4 Relevance

PAIR: a-b — (1-4)

[1] Could you recommend movies like Amelie?

[2] Do you like romance or comedy films? Amelie is in that genre.

[3] Yes I do. Movies like You've Got Mail and You Again, can you suggest others?

RESPONSE: you ve got mail is romantic comedy do you like that genre

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, rendering score 1 the least suitable.

PAIR: a-b — (1-4)

[1] Could you recommend movies like Amelie?

[2] Do you like romance or comedy films? Amelie is in that genre.

[3] Yes I do. Movies like You've Got Mail and You Again, can you suggest others?

RESPONSE: you ve got mail is romantic comedy do you like that genre

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, rendering score 1 the least suitable.

PAIR: a-b — (1-4)

[1] Could you recommend movies like Amelie?

[2] Do you like romance or comedy films? Amelie is in that genre.

[3] Yes I do. Movies like You've Got Mail and You Again, can you suggest others?

RESPONSE: you ve got mail is romantic comedy do you like that genre

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, rendering score 1 the least suitable.

PAIR: a-d — (1-4)

[1] Do you like Edgar Ramirez?

[2] I sure do. Edgar Ramirez starred in "Carlos", a thriller film co starring Martha Higareda. Are you interested in thriller films?

[3] Yes I enjoy watching thriller films. Can you tell me more about the move Carlos please?

[4] Sure. Ilich Ramirez Sanchez, aka 'Carlos,' is a Venezuelan-born Marxist revolutionary who aligns himself with the Palestinian cause and becomes the world's most notorious terrorist. He leads assaults on the meeting of OPEC ministers, taking them hostage and flying them from country to country seeking asylum, one of the most daring acts of terrorism in history.

[5] Wow this movie Carolos sounds very interesting. I will have to check it out over the weekend. Can you tell me when it was released thanks?

RESPONSE: you re welcome

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, rendering score 1 the least suitable.

PAIR: a-d — (1-4)

[1] Could you recommend something like the book Reflected in You?

[2] Do you enjoy the writing of this book's author Sylvia Day?

[3] Yes! What else has she written?

[4] Some of Sylvia Day's books include Seven Years to Sin and Reflected in You. Have you heard of any of these?

[5] What genres do some of he books fall under?

RESPONSE: afterburn is drama do you like dramas

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, since it is at least thematically relevant, thus rendering score 1 the least suitable

---

PAIR: a-d — (1-4)

[1] Could you recommend something like the book Reflected in You?

[2] Do you enjoy the writing of this book's author Sylvia Day?

[3] Yes! What else has she written?

[4] Some of Sylvia Day's books include Seven Years to Sin and Reflected in You. Have you heard of any of these?

[5] What genres do some of he books fall under?

RESPONSE: afterburn is drama

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, since it is at least thematically relevant, thus rendering score 1 the least suitable

---

PAIR: a-d — (1-4)

[1] Could you recommend something by Joseph Conrad?

[2] Yeah, it appears he wrote the book called The Secret Agent: A Simple Tale

[3] I see that is sort of an old book, was it written before his book the Tale of the Seaboard?

[4] It appears to be written in the year 1907, but I'm not exactly certain.

[5] I think he also wrote Heart of Darkness, can you tell me anything about that book?

RESPONSE: yes heart of darkness was written by joseph conrad

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: Evaluating the response as completely irrelevant is inaccurate, since it is at least thematically relevant, thus rendering score 1 the least suitable

---

C.1.5 Clarity

PAIR: a-b — (5-1)

[1] Could you recommend any works by Scott Westerfeld?

[2] Scott Westerfeld wrote The Uglies Trilogy, as well as Pretties ironically

[3] Interesting, can you give me a little more info on Uglies Trilogy? I might look to check that out this week.

[4] Sure, The Uglies Trilogy (Uglies, 1-3) was released in 2007. It is an interesting collection, definitely worth checking out

[5] Awesome, maybe I'll keep within the time period, any other books that released that year by any chance?

RESPONSE: yes it was released in 2007

DUBIOUS SCORE: **1**

ANNOTATOR: **b**

EXPLANATION: By no means, does the response display all 4 undesired properties relating to ambiguity and obscurity rendering score 1 the least suitable.

PAIR: e-a — (5-1)

[1] Do you like soccer? Have you heard of FC Bayern Munich?

[2] I do not like soccer. I think it's a German soccer team. Louis van Gaal is the coach. Do you like soccer?

[3] Not a fan either. But I like hockey. Have you heard of the Golden Knights?

RESPONSE: is golden knights las vegas professional ice hockey team i like chicago blackhawks they are schedule for game this thursday

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: By no means, does the response display all 4 undesired properties relating to ambiguity and obscurity. Given that the annotators we instructed to disregard any mistakes caused by the normalization applied on predictions as part of the automatic evaluation, score 1 is the least suitable.

PAIR: e-a — (5-1)

[1] Do you like soccer? Have you heard of FC Bayern Munich?

[2] I do not like soccer. I think it's a German soccer team. Louis van Gaal is the coach. Do you like soccer?

[3] Not a fan either. But I like hockey. Have you heard of the Golden Knights?

RESPONSE: i don t know much about him but i do know that he is football player do you like football

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: The response does not display all 4 undesired properties relating to ambiguity and obscurity. Given that the annotators we instructed to disregard any mistakes caused by the normalization applied on predictions as part of the automatic evaluation, score 1 is the least suitable.

PAIR: e-a — (5-1)

[1] I like Anne Fletcher

[2] Did you know she starred in The Proposal alongside Sandra Bullock?

RESPONSE: no i didn t know that what else has sandra bullock been in

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: The response does not display all 4 undesired properties relating to ambiguity and obscurity. Given that the annotators we instructed to disregard any mistakes caused by the normalization applied on predictions as part of the automatic evaluation, score 1 is the least suitable.

PAIR: c-a — (5-1)

[1] Do you know any books written by Chuck Palahniuk?

[2] Tell-All, Survivor and Snuff

[3] What year was 'Tell-All' released?

[4] In 2010

[5] Are there any other thrillers you could recommend?

RESPONSE: do you like thrillers

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: By no means does the response display all 4 undesired properties relating to ambiguity and obscurity. Given that the annotators we instructed to disregard any mistakes caused by the normalization applied on predictions as part of the automatic evaluation, score 1 is the least suitable.

PAIR: c-a — (5-1)

[1] I like the movie Looper. Could you recommend something similar?

[2] Sure. Looper stars Bruce Willis. Do you like his work?

[3] Yes! I am an action movie fan and I love Bruce Wilson

[4] Hudson Hawk is an action film also starring Bruce Willis. Have you seen it?

RESPONSE: bruce willis starred in hudson hawk hudson hawk is great movie do you like bruce willis movies

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: The response does not display all 4 undesired properties relating to ambiguity and obscurity. Given that the annotators we instructed to disregard any mistakes caused by the normalization applied on predictions as part of the automatic evaluation, score 1 is the least suitable.

PAIR: c-a — (5-1)

[1] I like peter pan

[2] Me, too! It was written by Michael Hague, who also wrote The Voyages of Doctor Dolittle. Have you read that?

[3] What genre is it

[4] It is speculative fiction in the fantasy genre of children's literature.

RESPONSE: it s speculative fiction do you like speculative fiction

DUBIOUS SCORE: **1**

ANNOTATOR: **a**

EXPLANATION: The response does not display all 4 undesired properties relating to ambiguity and obscurity. Given that the annotators we instructed to disregard any mistakes caused by the normalization applied on predictions as part of the automatic evaluation, score 1 is the least suitable.

C.1.6 Brevity

PAIR: e-a — (5-2)

[1] Do you like soccer? Have you heard of FC Bayern Munich?

[2] I do not like soccer. I think it's a German soccer team. Louis van Gaal is the coach. Do you like soccer?

[3] Not a fan either. But I like hockey. Have you heard of the Golden Knights?

RESPONSE: i don t know much about him but i do know he is football player do you like football

DUBIOUS SCORE: **2**

ANNOTATOR: **a**

EXPLANATION: Interpretation is not hindered, thus rendering score 2 as the least suitable.

PAIR: e-a — (5-2)

[1] Do you like soccer? Have you heard of FC Bayern Munich?

[2] I do not like soccer. I think it's a German soccer team. Louis van Gaal is the coach. Do you like soccer?

[3] Not a fan either. But I like hockey. Have you heard of the Golden Knights?

RESPONSE: i don t know much about him but i do know that he is football player do you like football

DUBIOUS SCORE: **2**

ANNOTATOR: **a**

EXPLANATION: Interpretation is not hindered, thus rendering score 2 as the least suitable.

PAIR: c-a — (5-2)

[1] I like the movie Looper. Could you recommend something similar?

[2] Sure. Looper stars Bruce Willis. Do you like his work?

[3] Yes! I am an action movie fan and I love Bruce Wilson

[4] Hudson Hawk is an action film also starring Bruce Willis. Have you seen it?

RESPONSE: no i have not who else stars in in

DUBIOUS SCORE: **2**

ANNOTATOR: **a**

EXPLANATION: Interpretation is not hindered, thus rendering score 2 as the least suitable.

PAIR: a-d — (5-2)

[1] Do you know anything about the author Pearl S. Buck?

[2] Yes, he wrote The Good Earth which is a Historical period drama. Have you read it?

[3] I haven't yet. Do you know what year that was released?

[4] It was released in 1931. Another book you might like that was released the same year was The Waves.

[5] I haven

[6] I see, do they sound interesting to you? Another book I can recommend is Women in Love.

[7] Thanks! I love the historical period dramas and will definitely check it out.

RESPONSE: no i haven t can you tell me more about women in love

DUBIOUS SCORE: **5**

ANNOTATOR: **a**

EXPLANATION: 'no I haven't' hinders interpretation, thus rendering score 5 as the least suitable.

C.1.7 Coherence

PAIR: d-e — (5-2)

[1] Do you have any knowledge of who Gil Junger is?

[2] Yes! Gil Junger is a director, some of Junger's work include Black Knight and 10 Things I Hate About You.

RESPONSE: can you tell me more about 10 things i hate about you

DUBIOUS SCORE: **2**

ANNOTATOR: **e**

EXPLANATION: Meaning interpretation is not hindered, considering that the annotators were instructed to disregard any mistakes caused by normalization. Therefore, score 2 is the least suitable.

PAIR: d-e — (5-2)

[1] Do you like the movie American Psycho? Someone recommended it to me, but I don't really know anything about it.

[2] Well, from what I know, the movie was written my Mary Harron. Would you like to know who acted it?

RESPONSE: no i haven t can you tell me more about mary harron

DUBIOUS SCORE: **5**

ANNOTATOR: **d**

EXPLANATION: Coherence is not perfect, thus rendering score 5 the least suitable.

PAIR: d-e — (5-2)

[1] Do you like the movie American Psycho? Someone recommended it to me, but I don't really know anything about it.

[2] Well, from what I know, the movie was written my Mary Harron. Would you like to know who acted it?

RESPONSE: no i haven t what else has mary harron written

DUBIOUS SCORE: **5**

ANNOTATOR: **d**

EXPLANATION: Coherence is not perfect, thus rendering score 5 the least suitable.

PAIR: d-e — (5-2)

[1] Can you tell me, does Simona Halep play tennis?

[2] I believe that Simona Halep is a Tennis athlete.

[3] Do you like Tennis? Who do you like? I like Victoria Azarenka

[4] I do not really watch tennis but I believe Victoria has lost matches to my faves Serena and Venus but who hasn't?

[5] Right, they have even lost to each other at some point. They are the best in the business and Venus even is part owner of the Miami Dolphins and Serena is a Fashion designer, did you know that?

RESPONSE: no i didn t know that i did know that serena is fashion designer too

DUBIOUS SCORE: **Subjective**

ANNOTATOR: -

EXPLANATION: The normalized response distorts interpretation. It can be either interpreted a one incoherent sentence, or as two coherent sentences: 'No I didn't know that. I did know that Serena is a fashion designer too.'

PAIR: e-a — (5-2)

[1] I like Anne Fletcher

[2] Did you know she starred in The Proposal alongside Sandra Bullock?

RESPONSE: i haven t seen proposal who else is in it

DUBIOUS SCORE: **2**

ANNOTATOR: **a**

EXPLANATION: Meaning interpretation is not hindered, considering that the annotators were instructed to disregard any mistakes caused by normalization. Therefore, score 2 is the least suitable.

PAIR: e-a — (5-2)

[1] I like Anne Fletcher

[2] Did you know she starred in The Proposal alongside Sandra Bullock?

RESPONSE: no i didn t know that what else has sandra bullock been in

DUBIOUS SCORE: **2**

ANNOTATOR: **a**

EXPLANATION: Meaning interpretation is not hindered, considering that the annotators were instructed to disregard any mistakes caused by normalization. Therefore, score 2 is the least suitable.

PAIR: c-d — (5-2)

[1] Who wrote It?

[2] The book "It" was written by Stephen King. The first time they made it into a movie, it starred Tim Curry from the Rocky Horror Picture Show. Do you like Horror movies and books?

[3] I love scary horror movies. Didn't Stephen King write The Stand as well?

[4] He did! And apparently he made an appearance in the film version of the book. he also wrote "The Shawshank Redemption" , but thats more of a prison drama.

[5] I loved Shawshank! Can you recommend more horror movies?

RESPONSE: yes he did he also starred in stand which is horror film do you like horror movies

DUBIOUS SCORE: **5**

ANNOTATOR: **c**

EXPLANATION: Coherence is not perfect, thus rendering score 5 the least suitable.

PAIR: c-d — (5-2)

[1] Who wrote It?

[2] The book "It" was written by Stephen King. The first time they made it into a movie, it starred Tim Curry from the Rocky Horror Picture Show. Do you like Horror movies and books?

[3] I love scary horror movies. Didn't Stephen King write The Stand as well?

[4] He did! And apparently he made an appearance in the film version of the book. he also wrote "The Shawshank Redemption" , but thats more of a prison drama.

[5] I loved Shawshank! Can you recommend more horror movies?

RESPONSE: sure stephen king also starred in last stand which is horror film do you like horror movies

DUBIOUS SCORE: **5**

ANNOTATOR: **c**

EXPLANATION: Coherence is not perfect, thus rendering score 5 the least suitable.

PAIR: c-d — (5-2)

[1] Who wrote It?

[2] The book ”It” was written by Stephen King. The first time they made it into a movie, it starred Tim Curry from the Rocky Horror Picture Show. Do you like Horror movies and books?

[3] I love scary horror movies. Didn’t Stephen King write The Stand as well?

[4] He did! And apparently he made an appearance in the film version of the book. he also wrote ”The Shawshank Redemption” , but thats more of a prison drama.

[5] I loved Shawshank! Can you recommend more horror movies?

RESPONSE: stand is great horror film it also stars rob lowe do you like him

DUBIOUS SCORE: **5**

ANNOTATOR: **c**

EXPLANATION: Coherence is not perfect, thus rendering score 5 the least suitable.

PAIR: c-a — (5-2)

[1] Could you recommend some similar movies to Public Enemies?

[2] Sure! Do you like the writers Ronan Bennett or Michael Mann at all?

[3] I really liked the actor Christian Bale and Johnny Depp in this movie but the story line was great too.

RESPONSE: yes johnny depp starred in that movie as well he also starred in x men 2 and x men 3 have you seen either of those

DUBIOUS SCORE: **2**

ANNOTATOR: **a**

EXPLANATION: Meaning interpretation is not hindered, considering that the annotators were instructed to disregard any mistakes caused by normalization. Therefore, score 2 is the least suitable.

PAIR: d-e — (5-2)

[1] Do you like Serena Williams?

[2] My grandmother loved to watch her and Venus Williams play.

[3] Yeah. Have you seen them play against each other? It was a very good match.

[4] I'm sure my grandmother has bit I haven't. Do you know which sister won?

[5] I think Serena won. Who do you like more? Serena or Venus?

RESPONSE: i like both of them but i think only venus is still playing

DUBIOUS SCORE: **2**

ANNOTATOR: **e**

EXPLANATION: Meaning interpretation is not hindered, considering that the annotators were instructed to disregard any mistakes caused by normalization. Therefore, score 2 is the least suitable.

PAIR: d-e — (5-2)

[1] Do you like Serena Williams?

[2] My grandmother loved to watch her and Venus Williams play.

[3] Yeah. Have you seen them play against each other? It was a very good match.

[4] I'm sure my grandmother has bit I haven't. Do you know which sister won?

[5] I think Serena won. Who do you like more? Serena or Venus?

RESPONSE: i like serena more i think she s great tennis player do you like tennis

DUBIOUS SCORE: **2**

ANNOTATOR: **e**

EXPLANATION: Meaning interpretation is not hindered, considering that the annotators were instructed to disregard any mistakes caused by normalization. Therefore, score 2 is the least suitable.

PAIR: d-e — (5-2)

[1] I loved Monty Python and the Holy Grail. Can you recommend something similar?

[2] That was a great movie. The Life of Brian was another Monty Python movie. Do you like John Cleese and Eric Idle?

[3] I love them. I also really like Terry Gilliam and Spike Milligan. So anything that is done by them is great.

[4] A Fish Called Wanda and Time Bandits are not Python movies but along the same lines.

[5] Time Bandits sounds good. Who's in that?

RESPONSE: he also starred in time bandits have you seen it

DUBIOUS SCORE: **5**

ANNOTATOR: **d**

EXPLANATION: Coherence is not perfect, thus rendering score 5 the least suitable.

C.2 Score distribution

Criteria	Score Pairs
Soundness	('5-5', 250), ('4-5', 15), ('5-4', 11), ('3-1', 4), ('3-4', 4), ('1-2', 3), ('4-2', 3), ('5-3', 2)
Conciseness	('5-5', 133), ('4-4', 38), ('5-4', 37), ('4-5', 27), ('3-3', 15), ('4-3', 11), ('3-5', 9), ('3-4', 8)
Completeness	('5-5', 196), ('1-1', 30), ('4-4', 26), ('5-4', 11), ('5-3', 10), ('4-5', 7), ('1-3', 5), ('1-2', 4)
Relevance	('5-5', 111), ('4-4', 58), ('5-4', 34), ('4-5', 27), ('3-5', 15), ('2-2', 10), ('2-4', 8), ('2-1', 8)
Clarity	('5-5', 223), ('5-4', 24), ('4-5', 10), ('4-4', 10), ('3-5', 9), ('5-3', 9), ('4-3', 5), ('3-4', 3)
Brevity	('5-5', 108), ('4-4', 59), ('5-4', 45), ('4-5', 24), ('4-3', 16), ('3-3', 14), ('5-3', 13), ('4-2', 6)
Coherence	('5-5', 154), ('4-4', 34), ('5-4', 28), ('4-5', 21), ('3-3', 11), ('3-4', 10), ('2-4', 7), ('5-3', 7)
Dialogue-act	('0-0', 205), ('1-0', 33), ('0-1', 32), ('0-2', 11), ('2-0', 7), ('2-2', 4), ('1-1', 4), ('2-1', 3)
Emotion	('0-0', 258), ('2-0', 20), ('1-0', 13), ('0-2', 5), ('0-1', 3), ('2-1', 1)
Communicative goal	('0-0', 164), ('2-2', 39), ('0-1', 24), ('1-0', 22), ('0-2', 20), ('2-0', 18), ('2-1', 8), ('1-2', 5)

Figure C.2: The scores assigned to the annotated responses for each criterion. The first value in the parentheses (e.g., 5-4) indicates the scores given to a response by a pair of annotators, The second value in the parentheses indicates the number of times this score pair has been assigned across the entire annotation set.

C.3 System-Oriented Inter-Annotator Agreement

Models	Soundness	Conciseness	Completeness	Relevance	Clarity	Brevity	Coherence	AVERAGE	Dialogue-act	Emotion	Communicative goal
Reference	0.742	**0.463**	**0.043**	**0.156**	0.174	0.371	**0.422**	**0.339**	0.076	-0.057	**-0.108**
Godel-Comb-Half	**0.826**	0.496	0.847	0.476	0.334	0.528	0.614	0.589	**-0.123**	-0.018	0.217
Godel-Un-Half	0.799	**0.712**	0.889	0.709	0.384	0.598	0.706	0.685	**0.142**	-0.029	0.574
Godel-Str-Shared	0.754	0.570	0.917	**0.715**	**0.725**	**0.340**	0.639	0.666	0.104	**-0.120**	0.377
Godel-Comb-Per-Half	0.774	0.674	0.911	0.525	**-0.04**	0.515	0.508	0.553	-0.112	**0.150**	0.379
Godel-Str-Per-Shared	**0.717**	0.592	**0.933**	0.664	0.688	**0.600**	**0.830**	**0.718**	0.120	-0.090	**0.590**

Figure C.3: Inter-annotator agreement computed for each system individually between the scores assigned to its responses in each annotation round. Agreement is measured using Krippendorf's *alpha* coefficient.

C.4 Pair-Oriented Inter-Annotator Agreement

Pairs	Soundness	Conciseness	Completeness	Relevance	Clarity	Brevity	Coherence	AVERAGE	Dialogue-act	Emotion	Communicative goal
a-b	**0.921**	0.773	0.952	0.771	0.603	**0.836**	0.623	**0.783**	**0.1**	**-0.304**	**0.1**
d-e	0.639	0.627	**0.954**	0.663	**-0.054**	0.434	0.746	0.573	**0.47**	-0.029	0.589
b-c	0.726	0.761	**1.000**	**0.886**	0.734	0.496	0.672	0.754	1.0	**1.0**	0.418
e-a	0.818	**0.256**	**0.750**	0.374	0.042	-0.002	**0.564**	**0.400**	**-0.198**	-0.046	**0.112**
c-d	0.405	0.556	0.810	0.722	0.457	**-0.088**	**0.772**	0.519	-0.154	-0.031	0.114
e-b	0.797	**0.931**	0.948	**-0.239**	0.0	0.461	0.758	0.522	-0.063	-0.063	0.672
c-a	**0.364**	0.510	0.915	0.771	0.146	0.303	0.595	0.515	-0.112	**1.0**	0.773
a-d	0.863	0.426	0.915	0.749	**0.748**	0.089	0.717	0.644	0.491	0.0	0.807

Figure C.4: Inter-annotator agreement computed between the scores within each of the 10 annotation pairs. Agreement is measured using Krippendorf's *alpha* coefficient.

Bibliography

M. Abdullah, A. Madain, and Y. Jararweh. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8, 2022. doi: 10.1109/SNAMS58071.2022.10062688.

D. Alon and J. Ko. Goemotions: a dataset for fine-grained emotion classification. *Google AI Blog*, 2021.

J. Auguste, F. Béchet, G. Damnati, and D. Charlet. Skip act vectors: integrating dialogue context into sentence embeddings. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 3–14. Springer, 2021.

S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

E. Bruni and R. Fernandez. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288, 2017.

P. Camara, M. C. Kloos, V. Kyrmanidi, A. Kluska, R. Terlou, and L. Krause. [re] dialsummeval-evaluation of automatic summarization evaluation metrics. In *ML Reproducibility Challenge 2022*, 2023.

C. Chakrabarti and G. F. Luger. A framework for simulating and evaluating artificial chatter bot conversations. In *The Twenty-Sixth International FLAIRS Conference*, 2013.

H. Chen, X. Liu, D. Yin, and J. Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

L. Chen, B. Lv, C. Wang, S. Zhu, B. Tan, and K. Yu. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7521–7528, 2020a.

N. Chen, Y. Wang, H. Jiang, D. Cai, Y. Li, Z. Chen, L. Wang, and J. Li. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters.

N. Chen, Y. Wang, H. Jiang, D. Cai, Y. Li, Z. Chen, L. Wang, and J. Li. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters, 2023.

X. Chen, C. Sun, J. Wang, S. Li, L. Si, M. Zhang, and G. Zhou. Aspect sentiment classification with document-level sentiment preference modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3677, 2020b.

M. Cheng, W. Wei, and C.-J. Hsieh. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335, 2019.

S. Choudhary, P. Srivastava, L. Ungar, and J. Sedoc. Domain aware neural dialog system. *arXiv preprint arXiv:1708.00897*, 2017.

D. Cole. The Chinese Room Argument. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023.

J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810, 2021.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

K.-P. Engelbrecht, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller. Modeling user satisfaction with hidden markov models. In *Proceedings of the SIGDIAL 2009 Conference*, pages 170–177, 2009.

G. Erkan. Language model-based document clustering using random walks. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 479–486, 2006.

X. Gao, Y. Zhang, M. Galley, C. Brockett, and B. Dolan. Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*, 2020.

Y. Ge, Z. Xiao, J. Diesner, H. Ji, K. Karahalios, and H. Sundaram. What should i ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. *arXiv preprint arXiv:2205.10977*, 2022.

M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A. B. Goldberg and X. Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing*, pages 45–52, 2006.

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen. Critic: Large language models can self-correct with tool-interactive critiquing, 2023.

H. P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

A. Haghighi, A. Y. Ng, and C. D. Manning. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 387–394, 2005.

W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

M. Hanna and O. Bojar. A fine-grained analysis of BERTScore. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online, Nov. 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wmt-1.59`.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

H. He, A. Balakrishnan, M. Eric, and P. Liang. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *arXiv preprint arXiv:1704.07130*, 2017.

B. Hixon, P. Clark, and H. Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 851–861, 2015.

W. Hu, Z. Chan, B. Liu, D. Zhao, J. Ma, and R. Yan. Gsn: A graph-structured network for multi-party dialogues. *arXiv preprint arXiv:1905.13637*, 2019.

Z. Ji, Z. Liu, N. Lee, T. Yu, B. Wilie, M. Zeng, and P. Fung. Reducing hallucination in open-domain dialogues with knowledge grounding. *arXiv preprint arXiv:2212.01588*, 2022.

Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

C. K. Joshi, F. Mi, and B. Faltings. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*, 2017.

P. Jwalapuram. Evaluating dialogs based on grice's maxims. In *Proceedings of the Student Research Workshop associated with RANLP*, pages 17–24, 2017.

P. Kapanipathi, V. Thost, S. S. Patel, S. Whitehead, I. Abdelaziz, A. Balakrishnan, M. Chang, K. Fadnis, C. Gunasekara, B. Makni, et al. Infusing knowledge into the textual entailment task using graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8074–8081, 2020.

H. Khayrallah and J. Sedoc. Measuring thei don't know'problem through the lens of gricean quantity. *arXiv preprint arXiv:2010.12786*, 2020.

K. Krippendorff. Content analysis: An introduction to its methodology (2 nd thousand oaks, 2004.

T. Lan, X.-L. Mao, W. Wei, X. Gao, and H. Huang. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–37, 2020.

S. Lappin. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*, pages 1–12, 2023.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

B. Liu and L. Wu. Graph neural networks in natural language processing. *Graph Neural Networks: Foundations, Frontiers, and Applications*, pages 463–481, 2022.

C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

X. Liu, X. You, X. Zhang, J. Wu, and P. Lv. Tensor graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8409–8416, 2020.

Y. Liu, S. Feng, D. Wang, K. Song, F. Ren, and Y. Zhang. A graph reasoning network for multi-turn response selection via customized pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13433–13442, 2021.

Y. Long, J. Wang, Z. Xu, Z. Wang, B. Wang, and Z. Wang. A knowledge enhanced generative conversational service agent. In *Proceedings of the 6th Dialog System Technology Challenges (DSTC6) Workshop*, 2017.

R. Lowe, N. Pow, I. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015.

R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*, 2017.

G. Marcus, E. Leivada, and E. Murphy. A sentence is worth a thousand pictures: Can large language models understand human language?, 2023.

S. Mehri and M. Eskenazi. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*, 2020a.

S. Mehri and M. Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*, 2020b.

S. Mehri, E. Razumovskaia, T. Zhao, and M. Eskenazi. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*, 2019.

R. Mihalcea. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 411–418, 2005.

T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

S. Moon, P. Shah, A. Kumar, and R. Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 845–854, 2019.

V. Nastase, R. Mihalcea, and D. R. Radev. A survey of graphs in natural language processing. *Natural Language Engineering*, 21(5):665–698, 2015.

OpenAI. Gpt-4 technical report, 2023.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

A. Parisi, Y. Zhao, and N. Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.

P. Parthasarathi and J. Pineau. Extending neural generative conversational model using external knowledge sources. *arXiv preprint arXiv:1809.05524*, 2018.

B. Peng, M. Galley, P. He, C. Brockett, L. Liden, E. Nouri, Z. Yu, B. Dolan, and J. Gao. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*, 2022.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

D. Ramage, A. N. Rafferty, and C. D. Manning. Random walks for text semantic similarity. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing (TextGraphs-4)*, pages 23–31, 2009.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information processing & management*, 33(2):193–207, 1997.

A. P. Saygin and I. Cicekli. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258, 2002.

I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

K. Shuster, J. Xu, M. Komeili, D. Ju, E. M. Smith, S. Roller, M. Ung, M. Chen, K. Arora, J. Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.

P. Sommerauer and A. Fokkens. Conceptual change and distributional semantic models: An exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, 2019.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. Dai, A. La, A. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, B. Orinion, C. Diao, C. Dour, C. Stinson, C. Argueta, C. F. Ramírez, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, D. Freeman, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, D. Schrader, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway, E. Pavlick, E. Rodola, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang,

E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed, F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascandolo, G. Mariani, G. Wang, G. Jaimovitch-López, G. Betz, G. Gur-Ari, H. Galijasevic, H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. Shevlin, H. Schütze, H. Yakura, H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocoń, J. Thompson, J. Wingfield, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller, J. U. Balis, J. Batchelder, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman, J. Guerr, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. D. Dhole, K. Gimpel, K. Omondi, K. Mathewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. O. Colón, L. Metz, L. K. Şenel, M. Bosma, M. Sap, M. ter Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. R. Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen, M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu, M. Ivanitskiy, M. Starritt, M. Strube, M. Swedrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Walker, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, M. V. T, N. Peng, N. A. Chi, N. Lee, N. G.-A. Krakover, N. Cameron, N. Roberts, N. Doiron, N. Martinez, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. Risco, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. LeBras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. Chi, R. Lee, R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, Shyamolima, Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin, S. Prasad, S. T. Piantadosi, S. M. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. Ramasesh, V. U. Prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, Y. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, and Z. Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Y.-L. Tuan, Y.-N. Chen, and H.-y. Lee. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. *arXiv preprint arXiv:1910.00610*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

H. Wang, Z. Lu, H. Li, and E. Chen. A dataset for research on short-text conversations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 935–945, 2013.

M. Wang, Z. Lu, H. Li, and Q. Liu. Syntax-based deep matching of short texts. *arXiv preprint arXiv:1503.02427*, 2015.

J. Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

T.-H. Wen, M. Gasic, D. Kim, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *arXiv preprint arXiv:1508.01755*, 2015.

J. Wu and H. Zhou. Augmenting topic aware knowledge-grounded conversations with dynamic built knowledge graphs. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 31–39, 2021.

L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long, et al. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2):119–328, 2023.

C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma. Topic aware neural response generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou. Hierarchical recurrent attention network for response generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

K. Xu, L. Wu, Z. Wang, Y. Feng, and V. Sheinin. Sql-to-text generation with graph-to-sequence model. *arXiv preprint arXiv:1809.05255*, 2018.

K. Yao, B. Peng, G. Zweig, and K.-F. Wong. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292*, 2016.

S. Yao, T. Wang, and X. Wan. Heterogeneous graph transformer for graph-to-sequence learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7145–7154, 2020.

T. Young, F. Xing, V. Pandelea, J. Ni, and E. Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629, 2022.

D. Yu and Z. Yu. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*, 2019.

D. Yu, K. Sun, C. Cardie, and D. Yu. Dialogue-based relation extraction. *arXiv preprint arXiv:2004.08056*, 2020.

H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120, 2002.

B. Zhang, Y. Zhang, R. Wang, Z. Li, and M. Zhang. Syntax-aware opinion role labeling with dependency graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3249–3258, 2020a.

N. Zhang, S. Deng, J. Li, X. Chen, W. Zhang, and H. Chen. Summarizing chinese medical answer with graph convolution networks and question-focused dual attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 15–24, 2020b.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL `https://aclanthology.org/P18-1205`.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

T. Zhao, R. Zhao, and M. Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.

W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33*, pages 338–349. Springer, 2011.

D. Zhou, X. Hu, and R. Wang. Neural topic modeling by incorporating document relationship graph. *arXiv preprint arXiv:2009.13972*, 2020.

H. Zhou, M. Huang, and X. Zhu. Context-aware natural language generation for spoken dialogue systems. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2032–2041, 2016.

H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.

W. Zhou, Q. Li, and C. Li. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. *arXiv preprint arXiv:2105.15171*, 2021.

Z. Zhou, D. Zhang, W. Xiao, N. Dingwall, X. Ma, A. O. Arnold, and B. Xiang. Learning dialogue representations from consecutive utterances. *arXiv preprint arXiv:2205.13568*, 2022.