Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Ellen Zhang Chang

Surrounding Dialogue Generation using Deep Learning with Adapters

Master's thesis in Computer Science
Supervisor: Ole Jakob Mengshoel
June 2022

NTNU
Norwegian University of
Science and Technology

Ellen Zhang Chang

Surrounding Dialogue Generation using Deep Learning with Adapters

Master's thesis in Computer Science
Supervisor: Ole Jakob Mengshoel
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Text generation is an active research area concerning many problems, including machine translation and response generation. However, these problems only concern the generation of following text, not preceding, to the best of our knowledge. We present the novel surrounding dialogue generation problem, which consists of adding preceding and following utterances to a snippet of a dialogue. Surrounding dialogue generation has many applications, including content creation for entertainment and educational purposes. For instance, human content creators can write a dialogue snippet, extend the snippet using a surrounding dialogue generation architecture, and make adjustments to the extended dialogue until satisfied. The adjusted dialogue can finally be released to the consumer as specialized content in the form of communication exercises or entertaining, story-driven games, for instance. This way, the content creation process is streamlined so a bigger audience of consumers, with different interests and needs, can receive relevant content.

We also present an approach for solving the surrounding dialogue generation problem. Specifically, we propose a deep learning architecture with adapters that extends dialogues by adding preceding and following utterances to a snippet of a dialogue in iterations. It uses the open-source pre-trained language model with state-of-the-art performance, the Generative Pre-trained Transformer 2. We also focus on developing an efficient solution, as recent trends within the Natural Language Processing field have brought concerns for the sustainability and scalability of language models. Through user studies and machine learning studies, we find that our architecture is beneficial as a creative tool for content creators. Within five minutes, the content creators can improve the extended dialogues to a satisfactory quality. Our adapter-based tuning approach is also more efficient in terms of training time, storage space, and memory usage during training, compared to fine-tuning.

Sammendrag

Tekstgenerering er et aktivt forskningsområde som angår mange problemer, inkludert maskinoversettelse og responsgenerering. Disse problemene gjelder imidlertid bare generering av etterfølgende tekst til en tekstsnutt, ikke tekst i forkant av en tekstsnutt, etter vår beste kunnskap. Vi presenterer det nye omliggende dialoggenereringsproblemet som består av å legge til uttalelser i forkant og etterkant av en dialogsnutt. Omliggende dialoggenerering har mange applikasjoner, inkludert innholdsskaping for underholdningsformål og pedagogiske formål. For eksempel kan menneskelige innholdsskapere skrive en dialogsnutt, utvide snutten ved å bruke en omliggende dialoggenereringsarkitektur og gjøre justeringer på den utvidede dialogen til de er fornøyde. Den justerte dialogen kan endelig publiseres til forbrukeren som spesialisert innhold i form av kommunikasjonsøvelser eller underholdende, historiedrevne spill, for eksempel. På denne måten blir innholdsskapingsprosessen effektivisert slik at flere forbrukere, med ulike interesser og behov, kan motta relevant innhold.

For å løse det omliggende dialoggenereringsproblemet forslår vi en dyplæringsarkitektur med adaptere som utvider dialogsnutter ved å legge til uttalelser i forkant og etterkant av dialogsnutten i iterasjoner. Den bruker den forhåndstrente språkmodellen med åpen kildekode og toppmoderne ytelse, Generative Pre-trained Transformer 2. Vi fokuserer også på å utvikle en effektiv løsning, ettersom nyere trender innen naturlig språkbehandlingsfeltet har skapt bekymringer for bærekraft og skalerbarhet av språkmodeller. Gjennom brukerstudier og maskinlæringsstudier finner vi at arkitekturen vår er gunstig som et kreativt verktøy for innholdsskapere. Innen fem minutter kan innholdsskaperne forbedre de utvidede dialogene til en tilfredsstillende kvalitet. Vår adapterbaserte fremgangsmåte er også mer effektiv når det gjelder treningstid, lagringsplass og minnebruk under trening, sammenlignet med finjustering.

Preface

This master's thesis was conducted as part of the course TDT4900 at the Norwegian University of Science and Technology, spring 2021.

I wish to thank my supervisor Ole Jakob Mengshoel for his valuable insights, suggestions for research articles, improvements to this work, enjoyable discussions, and for keeping my motivation throughout this project.

Finally, I wish to thank the people at the company I collaborated with for participating in my user studies, providing me with state-of-the-art technologies on natural language processing and on-field experience of its applications, and giving feedback on this work.

Ellen Zhang Chang
Trondheim, June 20, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter gives an insight into the background and motivation of this thesis in Section 1.1 and Section 1.2. The goal of the thesis and research questions are presented in Section 1.3. To reach the goal and answer the research questions, we apply the research method described in Section 1.4. A summary of our main contributions is given in Section 1.5. Finally, we give a description of the structure of the thesis in Section 1.6.

1.1 Context

In tandem with technology advancements, the amount of data stored and sent across the interwebs has increased. This has brought more attention to data-driven solutions for various tasks through artificial intelligence. This also applies to the Natural Language Processing (NLP) field, where Artificial Neural Networks (ANNs) are used to learn natural language from large amounts of textual data represented as word embeddings [9, 62, 71]. These ANNs are called language models and several state-of-the-art language models for solving NLP tasks are based on the (encoder-decoder) sequence-to-sequence structure [68]. A breakthrough is the Transformer [73], a model for machine translation that employs a sequence-to-sequence structure in a Deep Neural Network (DNN) architecture using only the attention mechanism. State-of-the-art language models, e.g., GPT-2 (Generative Pre-trained Transformer) [62], GPT-3 [9], and BERT (Bidirectional Encoder Representations from Transformers) [18], are based on the Transformer. These models have millions to billions of parameters, require days to years of training on strong GPUs and millions of dollars to maintain during training (including electricity and hardware maintenance costs) [18, 9, 62].

By adapting these large-scale language models, agents are now, with varying

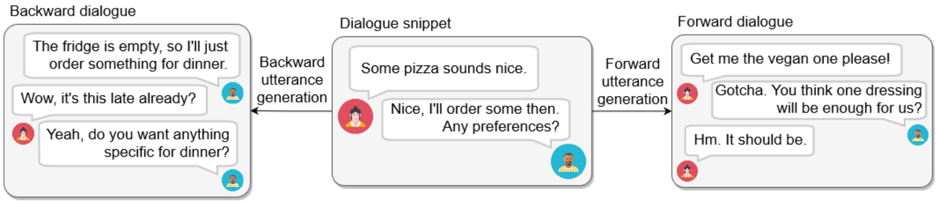


Figure 1.1: Given the snippet of a dialogue (in the middle) in a scenario of ordering a pizza, we generate surrounding dialogue through backward and forward utterance generation. Backward utterance generation (to the left) adds preceding utterances to the dialogue snippet. Forward utterance generation (to the right) adds following utterances to the dialogue snippet. The resulting dialogue is more meaningful compared to the dialogue snippet by itself.

success, capable of text generation [83, 21, 9, 73], sentiment analysis [29], and text classification [66] have been developed. Text generation tasks that have been researched consist of either transforming a sequence of text into a different sequence with the same meaning (e.g., machine translation [73], text style transfer [37], paraphrase generation [22], spelling correction [7], and document summarization [24]), or adding context to or a different meaning to it (e.g., distractor generation [23], story generation [21, 53], and utterance generation [71, 83, 31]). However, in the text generation tasks where sequences are extended (e.g., utterance generation), they are only extended in one direction: forward (i.e., adding to the end of a sequence). This also applies to dialogue systems, which are capable of conversing with a human being through utterance *response* generation (e.g., [75, 71, 31]). An area yet to be explored, to the best of our knowledge, is surrounding dialogue generation (see Figure 1.1), which includes generation of preceding utterances to a dialogue snippet.

1.2 Challenges

Surrounding dialogue generation has many applications, including the streamlining of the specialized content creation process. Examples of specialized content in the form of dialogues are communication exercises (e.g., for refugees to increase their language proficiency in areas of interest), and games for audiences with special requirements (e.g., limits in the ability to move). However, specialized content creation requires creativity, skills to write high-quality dialogues, and domain-specific knowledge. This also applies to the company we collaborate with. Taking into consideration the consumers, with different interests, it is unfeasible for a few content creators to satisfy all of them. However, machines can process data at a higher rate than humans, retain more information, and generate data faster. We hypothesize that surrounding dialogue generation can streamline the specialized content creation process and, thus, make more relevant content available for a broader audience.

Progress in Machine Learning (ML), Natural Language Processing (NLP), and Deep Learning (DL) have given us data-driven tools to assist humans in various tasks. Examples of this include grammatical error correctors [42] (which detects grammatical errors and suggests corrections) and machine translators [13] (which translates a word or phrase into a different language). They have increased the efficiency and quality of writing. However, these tools are computationally intensive to develop, as they are based on DL. The trend of state-of-the-art language models increase not only in performance, but also in size [62, 9]. Training these models leave an environmental footprint [71, 9]. Thus, a key challenge is the scalability, the computational resources required, and the sustainability of these data-driven tools for writing in natural language.

1.3 Goal and Research Questions

In the context of the challenges mentioned in Section 1.2, we define the main hypothesize of this thesis:

Hypothesis A simple, but efficient data-driven agent for surrounding dialogue generation can, when used in collaboration with human content creators, result in high-quality specialized content.

For our purposes, a simple data-driven agent is an artificial intelligence model that by itself is not capable of consistently outputting high-quality content. We hypothesize that the combination of human and machine effort can streamline the content creation process. Text is usually written in a forward manner, finishing a sentence word-by-word. However, the human mind does not always work in such a predictable manner. One can sometimes add a sentence to the beginning of, in the middle of, or at the end of a text, as new ideas pop into our minds. We hypothesize that getting drafts of what could be preceding and following utterances to a dialogue snippet can assist content creators in the creative process of writing dialogues (see Figure 1.2). Additionally, we consider the computational costs of state-of-the-art data-driven tools for writing dialogues.

We study in this thesis surrounding dialogue generation in English between two speakers, alternating in turns to say utterances, and formulate the following goal:

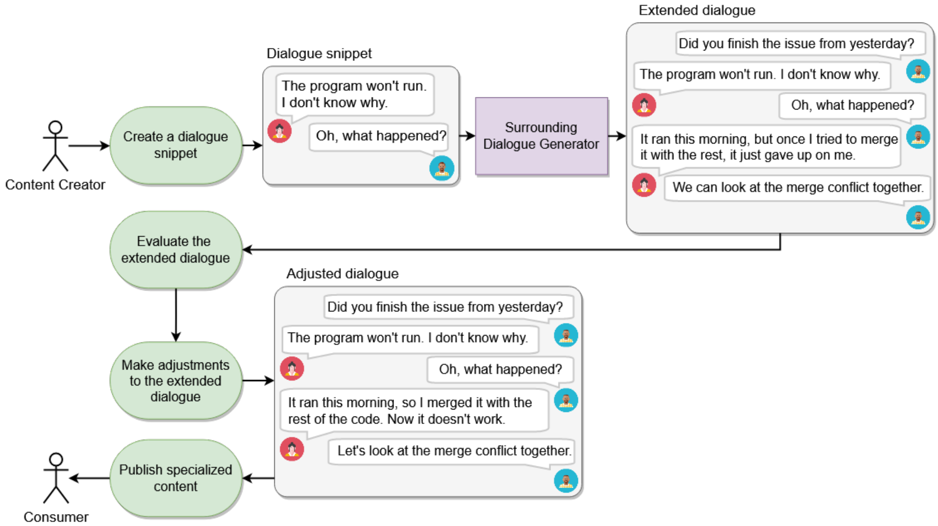


Figure 1.2: A content creator designs a dialogue snippet and sends it to a surrounding dialogue generator. The surrounding dialogue generator outputs a extended dialogue, which the content creator evaluates and makes adjustments to. The adjusted dialogue is then published and made available for the consumer.

Goal Develop an efficient data-driven architecture which generates, from a dialogue snippet including topic-specific terminology, surrounding topic-specific dialogue suitable as communication exercises for humans after adjustments made by an expert.

To reach the goal, the following research questions (RQs) are studied.

RQ1 How can large-scale, general, deep learning language models be adapted to topic-specific surrounding dialogue generation in a scalable way using adapters [6, 28]?

RQ2 How do data-driven approaches, for selecting generated preceding and following utterances to add to a dialogue snippet, align with human content creators' judgements?

RQ3 Do extended dialogues through data-driven approaches, after adjustments made by experts, result in high-quality dialogues appropriate as specialized content?

RQ4 How does the subjective and open-ended nature of content creation impact evaluations of extended and adjusted dialogues?

1.4 Research Method

This section described the research method we use to reach the goal of this master's thesis and answer the RQs. We continue our research from the specialization project [12], which concluded a literature study within the field of NLP with a proposed architecture for surrounding dialogue generation. Since the goal of this thesis is to develop a solution to a novel problem, we use a design-driven research method with collaboration with an external company in mind. Thus, the research is split into the following phases:

1. **Non-disclosure Agreement.** Since this master's thesis is done in collaboration with an external company, we discuss expectations and write and sign a non-disclosure agreement. This concluded with the company gaining ownership of the code, while we gain access to their technologies, get assistance in using their technologies, and can publish this thesis and papers related to it without any time restrictions. However, we cannot disclose the name of the company, nor their specific use case of our proposed architecture.

2. **Literature Study.** We continue the literature study from our specialization project [12] to stay up-to-date within the research field. This is because the literature study is the foundation for the thesis and our proposed architecture. Since the goal of this thesis is to solve a novel problem, we search for related work we can take inspiration from to design a solution grounded in theory and supported by results.

3. **Problem Formulation.** We define the problem, which is derived from the goal of this thesis. This clarifies the purpose of our proposed ML architecture and how it may be evaluated through experiments. This phase is done through discussions with content creation and technology experts at the company we collaborate with.

4. **Iterative Development.** Before diving into the development of a complex architecture, we first develop a simple version of the proposed architecture capable of generating surrounding dialogues from a snippet. Once the minimum viable product can solve the surrounding dialogue generation problem in a limited way, we develop improved versions of our proposed architecture in iterations.

5. **Evaluation.** We evaluate our proposed architecture through experiments in the form of user studies and studies of the efficiency of the ML methods. The goal of each experiment is to answer at least one of the RQs. In order to have an efficient and pleasant experience, we show drafts of the experiments to the company (and the participants from the company), make adjustments to the experimental methods in collaboration with the company, and conduct a small user test (on one of the participants) to evaluate the time costs of the experiments before conducting the real experiments. The final results are discussed with the supervisor of this thesis, the company, and the participants of the experiments.

6. **Conclusion.** We conclude this research with a summary of the process and our findings and propose future work. A final presentation is held for the company, and the code developed throughout the thesis is handed to the company.

Structured Literature Review Protocol We use the search engines Oria[1], Google Scholar[2] and ScienceDirect[3] to find research papers, with a preference for peer reviewed papers. Search words included: dialogue systems, narrative generation, dialogue generation, chatbot, chit-chat, attention mechanism, GPT-2, GPT-3, task-oriented dialogue systems, conversational dialogue systems, extending pre-trained models, transfer learning, evaluation methods for dialogue systems, text generation, and decoding methods. Semantic Scholar[4] is used to perform forward and backward citation tracking.

Questions for the literature research included:

- What makes a good dialogue?
- What natural language processing technologies are state-of-the-art?
- What dialogue systems are state-of-the-art?
- What transfer learning techniques for adapting large-scale language models are state-of-the-art?

We use no strict inclusion or evaluation criteria for literature due to the hole in the literature regarding this thesis goal. Instead, we require the papers to be relevant enough for the thesis goal or RQs and make a good enough impression to include them. The impression of the paper includes its number and strength

[1] https://oria.no/
[2] https://scholar.google.com/
[3] https://www.sciencedirect.com/
[4] https://www.semanticscholar.org/

of citations, the reputation of the journal it is published in, and the institution the research was conducted at. Papers with code are preferred. Discussions with authors and researchers are also taken into account. Throughout the semester, presentations of and discussions about various papers were held for the supervisor and his master's and doctoral education students, further adding to the impression of a paper.

1.5 Contributions

The main contributions of this thesis are discussed in Chapter 4 and Chapter 5. The following points provide a summary of the main contributions.

1. An architecture to generate a surrounding dialogue from a dialogue snippet.

2. An efficient transfer learning technique for forward and backward utterance generation with adapter-based tuning of a large language model.

3. An utterance scoring and selection method for adding preceding and following utterances to a dialogue snippet.

4. Results from a user study that suggest how human content creators can benefit from machine-generated surrounding dialogue.

5. Results from a questionnaire suggesting how a data-driven utterance scoring and selection method for adding preceding and following utterances to a dialogue snippet aligns with human content creators' judgement.

6. Results that show the efficiency of separate adapters for forward and backward utterance generation compared to fine-tuning.

1.6 Thesis Structure

The rest of the thesis is structured as follows.

Chapter 2: Background Theory In the next chapter, the background theory is presented, starting with state-of-the-art language models and their building blocks. Dialogue systems and evaluation challenges of them are introduced. Transfer learning techniques for NLP are also discussed. This chapter builds on previous work from our specialization project [12].

Chapter 3: Related Work Related work is presented along with how this thesis is placed in the literature. This chapter presents existing solutions to subtasks of the goal of the thesis and hypotheses on how existing work may be adapted to the thesis goal and its research questions.

Chapter 4: Proposed Architecture Our novel architecture, the BFD Generator, for generating surrounding dialogue is presented. This includes a description of its performance phase and learning phase. We also discuss our design choices and how our work differs from related work. The main use case of the BFD Generator for the company we collaborate with is as follows. The content creators in the company write a dialogue snippet which is relevant to their customers. The content creators make adjustments to the machine-generated extended dialogues to make them good enough as specialized communication content for their product. Finally, the specialized content is delivered to their customer.

Chapter 4: Experiments Four experiments are done to evaluate the proposed architecture. This includes user studies and machine learning studies. The results are presented and discussed in light of the research questions. Additionally, the limitations of the experiments are presented.

Chapter 5: Conclusion Finally, an evaluation and conclusion of this work are given by discussing the RQs and proposing future work.

Chapter 2

Background Theory

This chapter covers the background theory for dialogue systems, including the building blocks of state-of-the-art data-driven language models, evaluation methods, research areas, and challenges within the field.

2.1 Artificial Neural Network

Artificial Neural Networks (ANNs) are computing systems biologically inspired by the human nervous system [54]. ANN is a machine learning technique since it allows learning by example from a large amount of representative data that describes a physical phenomenon or decision process [63]. With ANNs, relationships between independent and dependent variables can be established without any mathematical assumption [63]. Another feature of ANNs is that they can extract subtle information and knowledge from representative data. ANNs have achieved excellent performance on difficult problems like speech recognition [5] and visual object recognition [38].

ANNs consist of layers of nodes (representing neurons), specifically, an input layer connected to an output layer through hidden layers [63]. A node may be connected to all or a subset of the nodes in the subsequent layer, simulating the synaptic connections of the brain. The input values of a node are multiplied by a weight on each connection, simulating the strengthening of neural pathways in the brain. A weighted signal entering a node simulates the electrical excitation of a nerve cell and, thus, the transference of information within the network. It is the adjustment of these connection weights in ANNs that emulate human learning [74].

Downstream Tasks ANNs solve downstream tasks. A node y is downstream of node x if and only if y uses information processed by x. This term is often used in NLP for tasks that use a pre-trained model or component, which are mostly ANNs.

Shortcomings A shortcoming of ANNs is their black-box nature. In most cases, it is not possible to find the assumptions under which each output node is most probable. The lack of explainability is especially a problem with Deep Neural Networks (DNNs), which are ANNs with multiple hidden layers. Gaining the trust of the people through explainability is important to justify the application of an agent. Thus, explainable artificial intelligence is an active research area [64, 26, 1].

Another problem with traditional neural networks is persistence. ANNs did not know how to use knowledge about previous information on which the new event is dependent, a key attribute of longer texts. This was the motivation behind Recurrent Neural Networks (RNNs). RNNs allow information to persist by introducing loops in the network. Long short-term memory and gated recurrent unit further address the vanishing gradient problem.

Despite the impressive advances in artificial intelligence and machine learning, it is important to mention the massive increase in computational cost, energy, and human resources for training these models [50]. DNNs typically require large datasets to learn enough general knowledge about a problem to solve it. Training a DNN from scratch for every single problem variant is too computationally heavy. Thus, when a DNN has been developed for a problem, finding out how to adapt it to another domain and transfer its knowledge is an important and active research area today [69, 47]. This is called transfer learning, a research area that is further discussed in Section 2.7.

Word and Sentence Embeddings In ANNs for NLP tasks, words are represented as vectors, so-called word embeddings. The idea is that words with similar meanings are closer in space. This makes it possible for machines, e.g., neural networks, to learn natural language from large amounts of data. It is possible to use pre-trained embedding spaces [56] or embed words as part of an architecture for a language model [73]. In some contexts, embedding of sentences or paragraphs may be interesting. The Universal Sentence Encoder [11] is an example of a data-driven model that is trained to encode sentences.

2.1.1 Sequence-to-Sequence

A sequence-to-sequence task, e.g., translation between natural languages, cannot always be sensibly encoded with vectors of fixed dimensionality. This is a

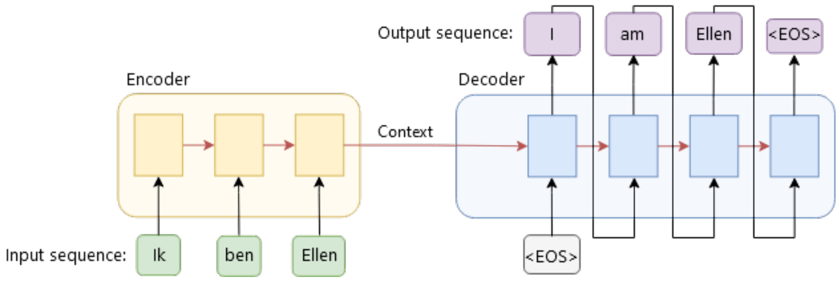


Figure 2.1: The seq2seq architecture showing translation of a Dutch sentence to English. The encoder sequentially transforms the input text to a context vector. The decoder uses the context vector and an end-of-sequence token (<EOS>) to generate the output text until an end-of-sequence token is generated.

potential problem since DNNs need big and labelled data of fixed dimensionality to solve difficult problems. This problem motivates the sequence-to-sequence (seq2seq) architecture of Sutskever et al. [68]. A seq2seq model takes a sequence of items as input and outputs another sequence of items, as seen in Figure 2.1. It is composed of an encoder and a decoder, which are usually RNNs [68]. The encoder maps the input into a latent representation (a context vector), which the decoder is conditioned on. Finally, the decoder generates the output sequentially. The seq2seq model often forms the basis for state-of-the-art language models.

However, the seq2seq model has some limitations. First, the encoder cannot capture the context of the dialogue by itself, as it only considers the current state and ignores all previous states [17]. Second, the seq2seq architecture is also prone to generating generic answers that follow the most common patterns of the training data [17].

2.1.2 Attention Mechanism

The attention mechanism is a deep learning technique inspired by the cognitive attention of humans and has been explored in cognitive science [8, 36] and artificial intelligence [4, 65]. When we are asked to count the number of faces in a picture, we do not pay an equal amount of attention to every part of the picture. We focus on the most important parts of the picture for the given task, namely the humans. This technique was a breakthrough in NLP [73, 61, 62, 9] and computer vision [38, 46] and is a state-of-the-art technique in both research fields.

The input to the attention mechanism consists of queries $\boldsymbol{q}$, keys $\boldsymbol{k}_i$, and values $\boldsymbol{v}_i$ for $i = 1, ..., N$, following the terminology of retrieval systems. To explain the attention mechanism in the context of NLP, a new word embedding with more context is queried for a token with existing word embedding $\boldsymbol{v_i}$. The existing database with the word embeddings of all tokens (the keys) is extracted for computing the dot product of the keys with the queried token. The results, called the scores, are normalized and named weights α_i. The existing database with the word embeddings of all tokens, called the values, are extracted and taken the dot product of with the weights. The resulting output is the new word embedding of $\boldsymbol{v_i}$, namely $\boldsymbol{y_i}$.

In the context of the encoder-decoder structure, the attention mechanism uses the hidden states from all of the layers in the encoder and decoder as input, compared to traditional RNNs which use the hidden states from the final layer only. To create the queries, keys, and values, a linear transformation of the aforementioned states is done, as seen in Equation 2.1:

$$\boldsymbol{q} = W^{(q)}\hat{\boldsymbol{q}} \qquad \boldsymbol{k}_i = W^{(k)}\hat{\boldsymbol{k}}_i \qquad \boldsymbol{v}_i = W^{(v)}\hat{\boldsymbol{v}}_i \tag{2.1}$$

The output $\boldsymbol{y}$ of the attention mechanism is the weighted sum of the values $\boldsymbol{v}_i$ for $i = 1, ..., N$, where the weights α_i are computed by some function f (see Equation 2.2). Each weight indicates which parts of the input values should be paid more attention to. The weights are normalized to avoid scaling the values. The weights are often represented by a so-called attention matrix.

$$\boldsymbol{y} = \sum_{i=1}^{N} \alpha_i \boldsymbol{v}_i \qquad \text{where } \alpha_i = f(\boldsymbol{q}, \boldsymbol{k}_i) \text{ and } \sum_{i=1}^{N} \alpha_i = 1 \tag{2.2}$$

A common approach to computing the weights is as seen in Equation 2.3:

$$\{\alpha_i\}_{i=1}^{N} = \text{softmax}(\{\tanh(\boldsymbol{q}^T \boldsymbol{k}_i)\}_{i=1}^{N}) \tag{2.3}$$

The hyperbolic tangent function gathers the stronger beliefs together and separates the weaker. Softmax heightens higher scores and depresses lower scores. This allows the model to be more confident about what to attend to.

Self-Attention Knowing how words relate to each other in a text is important to solve NLP tasks. This can be done with the attention mechanism by having the keys, queries, and values come from the same source. This is called self-attention. A good feature of self-attention is that proximity and order of the input do not influence the word embedding. Self-attention has been successful in a variety of tasks like translation [73] and reading comprehension [15].

Multi-Head Attention The word ambiguity problem is that words can have different meanings in different contexts and should be attended to in different ways depending on the context. Vaswani et al. [73] stumbled upon this problem when researching data-driven translation and extended self-attention with the proposal of multi-head attention. Multi-head attention creates multiple query, key, and value combinations, each $(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v})$-combination leading to a separate attention matrix.

2.2 State-Of-The-Art Language Models

A language model is a statistical model that attempts to understand natural language through observation, outputting probabilities for the next word given some context. Language models have many applications including document classification [2], chatbots [77], sentiment analysis [29], and opinion mining [1]. All of these used pre-trained language models to solve downstream tasks. In 2017, Vaswani et al. [73] showed that their neural network architecture Transformer, using only the attention mechanism, outperformed the state-of-the-art (SOTA) solutions. This was a breakthrough in NLP, and many other architectures based on the Transformer were developed and make up the current SOTA architectures [9, 18]. This section describes various SOTA architectures for text generation.

2.2.1 Transformer

The Transformer is a DNN architecture for data-driven language translation [73]. It follows the encoder-decoder structure of seq2seq (see section 2.1.1) and translation is done in similar fashion. Thus, the sentence is translated sequentially until an end-of-sequence token is generated. A sentence is translated by choosing the highest probable word in the output of the decoder and feeding it back into the decoder. It consists of a stack of six encoders and and a stack of six decoders (see Figure 2.2).

In Figure 2.3, the layers of the encoder at the top of the encoder stack (in Figure 2.2) are shown. The encoder takes the text to be translated as input. The decoder takes the target translation of the text. The text is first represented as word embeddings, then a sinusoidal positional encoding is added as additional context. This context attempts to capture the different meanings words can have depending on their position in a sentence.

The encoder consists of a multi-head attention layer and a feed-forward layer. The decoder consists of a masked multi-head attention layer, multi-head attention, and a feed-forward layer. After each of the layers, layer normalization is performed further stabilizing the results. This normalizes the vectors across each feature instead of the sample. The vectors from the decoder are passed through

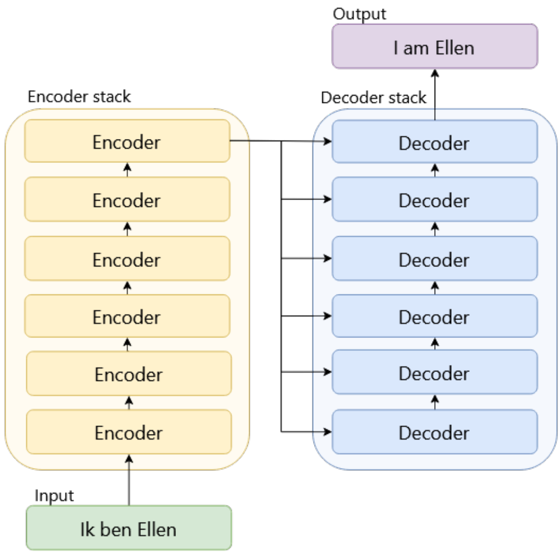


Figure 2.2: The Transformer's encoder stack (to the left) and decoder stack (to the right). The encoder stack consists of six encoders, and the decoder stack consists of six decoder. By viewing the encoder stack as an encoder, and the decoder stack as a decoder, the Transformer is similar to the seq2seq architecture (see Figure 2.1).

a linear layer that expands the dimensions to the vocabulary size of the language to be translated. Finally, the softmax layer transforms it to a probability distribution. This represents the agent's belief of what the next word can be when translating. Words can be sent in sequentially to translate a whole sentence.

The encoder first passes the word embeddings through multi-head attention, comparing the input with itself. This captures the word ambiguity problem (mentioned in Section 2.1.2). Then, it is fed through a position-wise fully-connected feed-forward layer to make it easier to digest by the decoder.

The decoder first passes the target translation input through masked multi-head attention. In masked multi-head attention, the right-context is masked. This prevents target leakage, as the model should not know what the next words are to predict the next word. Afterwards, it is passed through multi-head attention with the vectors from the encoder. This determines how related each word is to each other, encapsulating the interactions between the different languages. Finally, it is passed through a feed-forward layer like the encoder. Causal lan-

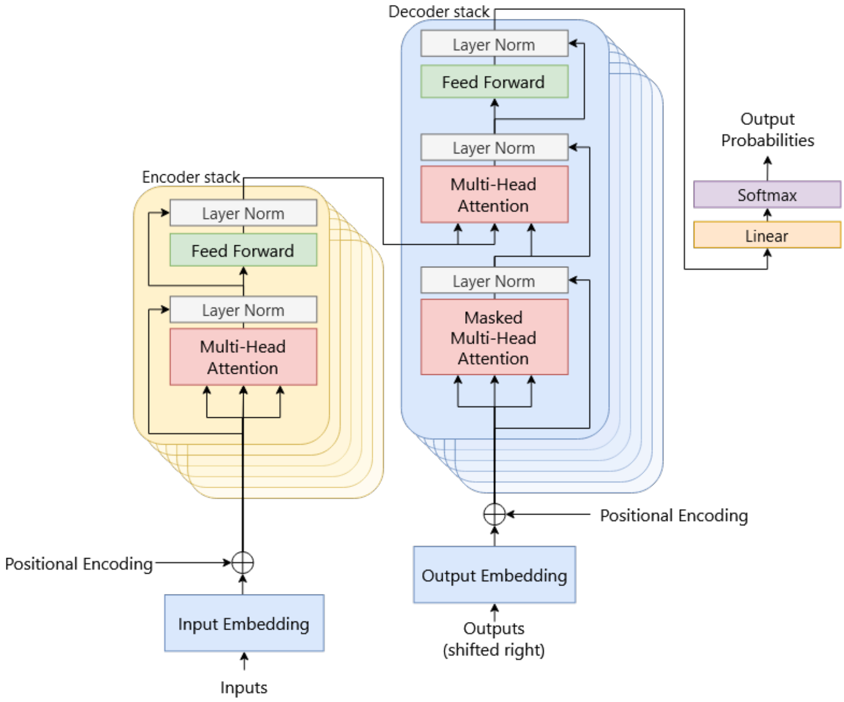


Figure 2.3: An overview of the Transformer architecture and the layers of the encoder and decoder.

Source: Vaswani et al. [73]

guage models use this decoding technique to predict the next word, given a word sequence (see Section 2.2.2).

2.2.2 Causal Language Models

Causal language models use the left context (text seen so far) to predict the next word, capable of text generation with varying success. Let x^t be a single training sequence of which we wish to model the joint probability over it. Given training sequences of the form $x = (x_1, ..., x_n)$, where each x_i comes from a fixed set of symbols, the goal of the language modeling is to learn $p(x)$. This distribution can be factorized using the chain rule of probability theory. The neural network is then trained with the parameters θ to minimize the cross-entropy loss or negative log-likelihood $\mathcal{L}$ over a dataset $D = \{x^1, ..., x^{|D|}\}$ where sequence x^t has length n_t. See Equation 2.4.

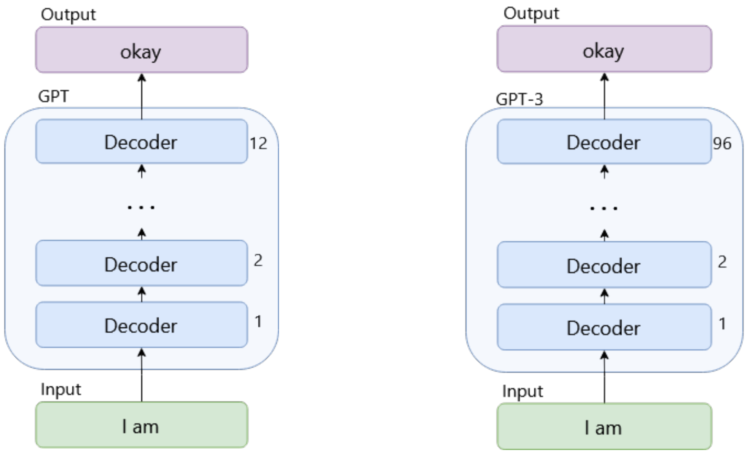
$$p(x) = \prod_{i=1}^{n} p(x_i|x_{<i}) \qquad \mathcal{L}(D) = -\sum_{t=1}^{|D|}\sum_{i=1}^{n_t} \log p_\theta(x_i^t|x_{<i}^t) \tag{2.4}$$

GPT (Generative Pre-trained Transformer) [61] is a causal language model, succeeded by GPT-2 [62] and the SOTA model GPT-3 [9]. GPT is a decoder-only Transformer-based model consisting of 12 decoder layers (see Figure 2.4a). GPT-3 further extends and improves the original Transformer model by pre-training on a larger corpus and adding more layers (see Figure 2.4b). The performance increase of GPT-3 is highly correlated to the increase in the number of parameters and training time. The largest GPT-3 model has 175 billion parameters [9]. This is two orders of magnitude more than the largest GPT-2 model. Even though the larger models have achieved better performance on many NLP tasks, only the smaller models are feasible for this project[1]. It has been seen that the larger model, the closer they are to being few-shot learners [9]. Few-shot learners only need a few training examples to perform well. Thus, a large dataset is most likely needed for this project.

GPT-Neo and GPT-J were developed as open-source alternatives to the commercialized GPT-3 and have achieved similar performance to the small GPT-3 models (125M, 1.3B, 2.7B, and 6B parameters).[2] It is important to note that GPT-Neo and GPT-J only adopt the architecture of GPT-3. To the best of our knowledge, no peer-reviewed research papers of GPT-Neo and GPT-J have been released as of spring 2021.

[1]We conducted a project as part of the NTNU course TDT13 Advanced Text Analytics and Language Understanding, fall 2021. The training was done on NTNU's high-performance computer. One of the biggest challenges was memory issues due to input and model size.

[2]https://nlpcloud.io/gpt-3-open-source-alternatives-gpt-j-gpt-neo.html



(a) GPT (b) GPT-3

Figure 2.4: GPT and GPT-3 both consist of layers of Transformer decoders; the decoder is described in Section 2.2.1.

2.2.3 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a SOTA pre-trained language model based of the Transformer-encoder [18]. Much like GPT, it consists of many layers of Transformers but uses the Transformer-encoders instead of the decoders. BERT also uses the context bidirectionally to predict, unlike the causal language models. This is why BERT excels in downstream tasks like Named-Entity Recognition (NER) and classification. Various BERT models have been developed for different tasks and languages by training the model on different data [39, 59] and making modifications or extensions to the architecture [69].

2.3 Markov Assumption

The Markov assumption is based on the idea that the next state can be predicted using a finite (and typically small) number of preceding states [25]. When applied to NLP, the probability $P(w_i|w_1, ..., w_{i-1})$ of the next word w_i to a sequence of words $(w_1, ..., w_{i-1})$ can be predicted using only the last j words $(w_{i-j}, ..., w_{i-1})$, i.e.:

$$P(w_i|w_1, ..., w_{i-1}) = \begin{cases} P(w_i|w_{i-1}) & \text{if } j = 0 \\ P(w_i|w_{i-j}, ..., w_{i-1}) & \text{if } 0 < j < i \end{cases} \tag{2.5}$$

This is called a jth order Markov chain. Most language models [62] and NLP architectures [31, 75] are based on the Markov assumption. It also allows for shorter sequences to be used during the training and performance phase of language models, making NLP problems more feasible.

2.4 Decoding Methods

Language models only output a probability distribution for the next token given a context. To transform the output into natural language, a decoding method is needed. The term *token* is often used instead of a word, as there are many ways to split phrases into words (e.g., "don't" or "do" and "not"). Top-k sampling, beam search, and nucleus sampling are among the most common decoding methods for text generation [80]. We now briefly present some decoding methods.

Greedy Search A simple, but naïve heuristic for decoding is greedy search, where the token with the highest estimated likelihood is picked. This approach is prone to generating boring, non-natural and redundant text, as human-written text is typically more surprising [30]. However, greedy search has shown promising performance when precision is the goal, e.g., in task-oriented dialogue systems [31]. It is possible to augment greedy search steps with noise to avoid local optima, which has been done in local search literature [49].

Beam Search Beam search improves greedy search: at every position in a sequence, it explores the b most probable tokens and continues exploring in such a fashion until it comes across an end-of-sequence token. When b sequences (ending with an end-of-sequence token) have been generated, their probabilities are computed. Finally, the sequence with the highest probability is chosen. Beam search is an effective strategy to sample sufficiently likely sequences from token probability distributions [68]. This is suitable for stricter tasks like text translation between languages. However, beam search is bad for open-ended tasks since it does not generate diverse text [40]. Additionally, a fundamental concern with beam search is the search space size.

Top-k Sampling By introducing noise to greedy search, more engaging and natural text can be generated. In top-k sampling, the top k most probable tokens are randomly sampled among. A drawback of this approach is that the

k-parameter does not consider whether the probability distribution is narrow or broad. This opens up for an over-representation of low-probability sequences, which can make the generated text trail off the context [30]. Top-k sampling also addresses the concern of search space size by pruning away less probable tokens.

Top-p Sampling To cut the over-representation of low-probability sequences down, top-p sampling (also called nucleus sampling) was proposed [30]. Top-p sampling computes the cumulative distribution from most probable to least until it passes the predefined value $p \in (0, 1)$. The left-over tokens are disregarded. Essentially, when the model has broad distribution for a context, more tokens are taken into consideration. When the model has a narrow distribution, fewer tokens are sampled. This suppresses the over-representation of low-probability sequences, minimizing the chance of trailing off. Thus, nucleus sampling can be suitable for generating engaging and interesting text [30, 49]. Top-p sampling also addresses the concern of search space size by pruning away less probable tokens.

2.5 Dialogue Systems

A dialogue system is a computer system that can have a dialogue with a user, e.g. response generation systems. In such systems, a dialogue is usually structured such that each speaker takes turns saying utterances (e.g., [71, 67, 75]), as seen in Figure 2.5. Two consecutive turns between different speakers make up an exchange. Multiple exchanges are considered a dialogue. After having a conversation with someone, only some snippets of the dialogue may stand out as memorable. From those, it is possible to speculate about the rest of the dialogue. Another interpretation of dialogues is that each turn or utterance is seen as an action (e.g., [78, 31]). A speaker takes an action by uttering, which changes the environment and the actions the speakers can take. Depending on their purpose, dialogue systems can be classified into three broad categories: task-oriented, conversational, or question-answering [17]. This work focuses on task-oriented and conversational dialogue systems and presents them in the following sections.

Task-oriented Dialogue Systems The goal of a task-oriented dialogue system is to efficiently help a user achieve his or her goal, which is usually clearly defined and measurable. Other characteristics of a task-oriented dialogue system are its structured behavior, specialization within a closed domain, and focus on efficiency [17]. Usually, the dialogue system initiates the dialogue with a user. Applications include technical support [44] and recommendation systems [14].

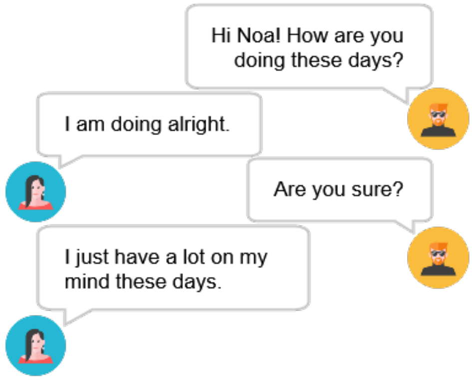


Figure 2.5: A dialogue between two speakers (indicated by the blue and yellow avatars), each taking turns saying utterances (in the speech balloons) while making exchanges.

Traditionally, task-oriented dialogue systems are designed as a pipeline consisting of a dialogue manager, a natural language understanding unit (NLU), a dialogue state tracker (DST), and a natural language generation unit (NLG) [17]. The dialogue manager is the core component and decides what dialogue action the system should take. The NLU extracts information from the user's utterance and identifies the corresponding dialogue act. The DST infers the current state of the dialogue. Finally, the output to the user is generated through the NLG unit. The problem with this traditional approach is that each unit is trained and supervised independently, making the pipeline vulnerable to error propagation across the components [43]. On the other hand, divide-and-conquer is a proven engineering technique, where the problem is divided into smaller tasks, and each of them are conquered separately.

A problem with this traditional pipeline is that each unit is trained and supervised independently, making the pipeline vulnerable to error propagation across the units [43]. Consequently and following Peng et al. [55], recent trends have moved toward the unification of the units (e.g., [81, 31]). SimpleTOD (Simple Task-Oriented Dialogue) [31] is a simple approach with state-of-the-art performance, which uses GPT-2 [62] to generate responses to task-oriented dialogue. SimpleTOD solves the sub-tasks of the different units in a unified way through multi-task maximum likelihood training. It enables modelling of the inherent dependencies between the sub-tasks of task-oriented dialogue, by optimizing for all tasks in an end-to-end manner.

Conversational Dialogue Systems A conversational dialogue system is often designed to keep an engaging conversation with the user [17]. The dialogues are usually unstructured and open-domain, with context and variability in utterances being important features. An interesting and engaging dialogue is kept if there is a satisfying degree of variation in topic and language. However, the context should not fluctuate too much, otherwise, the attentiveness of the user may be lost. The two main approaches to building dialogue systems are rule-based and data-driven (e.g., [79, 75]). A data-driven dialogue system typically uses either utterance classification or utterance generation.[3] Example conversational dialogue agent designs include chatbots with personality [19] and agents mimicking movie characters [79]. In the Second Conversational Intelligence Challenge, the conversational dataset PERSONA-CHAT was introduced. Wolf et al. [75] used dialogue-only data to fine-tune a dialogue system. It proved to be a state-of-the-art conversational dialogue system and won the automatic metrics track.

Hybrid and General Dialogue Systems Sun et al. [67] propose a task-oriented dialogue system enhanced with chit-chat. Their system consists of two language models and a switch module that decides their interactions depending on the context. The integration of conversational dialogue elements led to a more natural and engaging dialogue. LaMDA (Language Models for Dialog Applications) is a pre-trained deep learning language model [71]; it is closely related to the LaMDA metrics discussed in the paragraph about evaluation of dialogue systems. LaMDA is costly, but achieves great results with the ability to both generate and rank its generated responses. When fine-tuned on specific metrics, LaMDA can achieve near-human performance on sensibleness, specificity and interestingness. While fine-tuning on a small set of safety and groundedness labelled data showed increased performance, LaMDA's gap to human performance is still significant.

2.6 Evaluation of Dialogue Systems

Research on reliable, cheap, generalized automatic metrics for what makes a good dialogue or response is still an active research area [17]. Today, no automatic metrics can compare to human judgements. Additionally, most evaluation methods are highly correlated to specific characteristics of the system in mind [17]. While human judgements are often used in dialogue evaluations, they are expensive and not always reliable. The value of human judgements is further seen in LaMDA [71]. For generation, labelling and evaluation of dialogue training data, human

[3]Some dialogue systems do not generate utterances, but consider it a classification problem and pick from a selection of human-written utterances. Other systems generate utterances by looking at the context (either the dialogue or elements of the dialogue). This dichotomy applies both to conversational and task-oriented systems.

judgement is used. Typically, crowdworkers label responses given dialogue contexts, and rates them using, in the case of LaMDA, the following metrics: sensibleness [3], specificity [3], interestingness, safety, groundedness, informativeness, citation accuracy, helpfulness, and role consistency (see Section 2.6.1).

2.6.1 Gold-Standard Human Evaluation

The gold standard for evaluating dialogues is human evaluation, where humans assess the quality of a generated utterances or dialogues with guidelines like the following metrics.

1. **Sensibleness.** How well the utterances make sense in the dialogue context [3].
2. **Specificity.** How specific the utterances in the dialogue is [3].
3. **Interestingness.** How interesting the dialogue is [71].
4. **Informativeness.** The percentage of responses that carry information on the external world that can be supported by the other utterances in the dialogue [71].
5. **Groundedness.** The percentage of utterances that carry information on the external world that can be supported by external sources [71].
6. **Teachability.** Since we focus on communication, we propose a new metric, teachability, which uses the Common European Framework of Reference for Languages (CEFR). CEFR is an international standard for describing language proficiency, and organizes language proficiency in six levels, A1 to C2 [52]. Teachability measures how well the specific dialogue can be understood at a given language proficiency level based on CEFR scoring.

We use these metrics in some of our experiments in Chapter 5.

2.6.2 Automatic Metrics

In this section, automatic metrics that has been used in related work for evaluating language models, utterances or dialogues are described.

Cosine Similarity It is possible to compare similarity between words when they are represented as word embeddings through cosine similarity. As mentioned in Section 2.1, words with similar meanings are placed closer to each other in the word embedding space. By computing their distance in space, it is possible to compare words. The cosine similarity score *sim* between two word embeddings, $\boldsymbol{x}$ and $\boldsymbol{y}$, is computed as in Equation 2.6.

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{2.6}$$

Cosine similarity has been used as a measure of the similarity between the ground truth and the generated response in conversational dialogue systems [45], though its reliability is highly correlated to the reliability of the word or sentence embedding space. We use cosine similarity in our proposed architecture (see Equation 4.5).

Perplexity A measurement of how well a probability model predicts a sequence is perplexity. Perplexity of a sequence $W = (w_1, ..., w_N)$ is defined as:

$$\text{PPL}(D) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1, ..., w_{i-1})}}. \tag{2.7}$$

2.7 Transfer Learning

Transfer learning is a research area within machine learning and concerns how to transfer knowledge into a model. Training a data-driven language model is computationally heavy and time-consuming. Thus, adapting a pre-trained model without training it from scratch is an important research area. Another problem with training a language model from scratch is that domain-specific datasets are often sparse, making the models prone to overfitting. A reason for interest in transfer learning in NLP is that domain-specific language may not be represented by a general language model due to a change in sentence styles, formality, intent, and so on [69]. A word can also have different meanings in different domains [70]. This section presents two approaches to adapt a large-scale DNN language model to a specific task.

Fine-tuning Using a pre-trained model as initialization for training a new model on problem-specific data is called fine-tuning. This is a common transfer learning approach for DNN large-scale language models (e.g., [75, 71, 69]). There are various fine-tuning strategies, including training the whole model [75] or freezing of some pre-trained weights, usually whole layers [69]. Other aspects of fine-tuning include input representation and loss function(s). Many dialogue systems have been developed by fine-tuning a large-scale general language model that is capable of text generation (though not in the writing style of dialogues) (e.g., [31, 75]). For instance, Wolf et al. [75] develop a SOTA conversational dialogue system with personality by fine-tuning GPT-2 on the Persona-Chat corpus [82]. Fine-tuning of large-scale open-source language models is accessible through the `transformers` library [76].

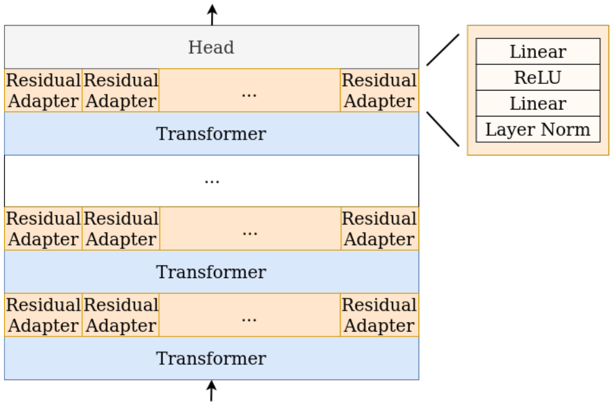


Figure 2.6: An example of how adapter-based tuning is applied, here in the AdapterCL architecture [47] (described in Section 3.4). Small neural networks called adapters are placed on top of each Transformer-layer in the GPT-2 model to adapt the language model without changing the original model's weights.

Source: Madotto et al. [47]

Adapter-based Tuning Adapter-based tuning is a parameter-efficient alternative to fine-tuning [32]. Unlike fine-tuning, adapter-based tuning freezes all layers of the pre-trained model. Separate, simple and small trainable neural networks, called adapters, are added between the layers of the pre-trained model, which steer the output distribution of a pre-trained model without modifying its original weights (see Figure 2.6) [47]. Adapter-based tuning has achieved comparable results to fine-tuning, while being more parameter efficient and allowing a high degree of parameter sharing [32]. Adapter-based tuning also outperforms fine-tuning on low-resource and cross-lingual tasks and is more robust to overfitting and less sensitive to changes in learning rates [28]. Adapter-based tuning has been used for a variety of NLP tasks like emotion recognition in dialogues [60] and machine translation [58]. Adapter-based tuning of large-scale open-source language models is accessible through the `adapter-transformers` library [57].

2.8 Noise in Dialogue Datasets

There are many ways noise can appear in dialogue datasets. An example is the task-oriented dialogue dataset MultiWOZ 2.1 of which Hosseini-Asl et al. identified four types of noisy labels in the human-annotated belief states. This includes misspellings of belief state values according to context information, non-labelled belief state when context provides sufficient information, labelled belief state when context lacks necessary information, and user-provided multiple options when the context does not provide sufficient information to determine the true belief state [31].

Other types of noises in dialogues can be misspellings, elongated words, a mix of languages, abbreviations, and many more. Depending on the handling of the corpus' language, features can be misrepresented as noise. For instance, social media has its own languages with emoticons, hashtags, images, etc., to convey messages. A lack of understanding, leading to misrepresentation of the data, can introduce unnecessary noise to the data.

Delexicalization Delexicalization is a common noise-reduction technique to improve the performance of NLP applications by removing language-specific phrases with language-agnostic terms [10, 31]. An example is seen in Figure 2.7. Performance is improved by generalizing over-specific training examples, making it easier for the model to see patterns and learn. Lexicalization of the machine-generated texts must be handled to regain their meaningfulness.

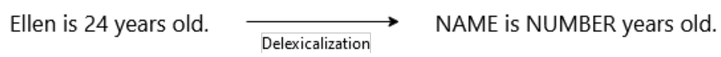


Figure 2.7: An example of delexicalization, where the left sentence is transformed to the right sentence.

Removal of stopwords Frequence words that do not add meaning to a sentence are called stopwords. Removal of stopwords is a common pre-processing technique in various NLP tasks, including dialogue generation [16], to decrease run-time and dilution of the meaning of text. Examples of stopwords are: “a”, “the”, and “is”.

Chapter 3

Related Work

Our proposed architecture adapts existing transfer learning techniques to reach the goal. It decomposes the problem, similar to task-oriented dialogue systems, but is more of a conversational dialogue system, given the creative nature of content creation. In this chapter, we present related work and how our work builds upon it and differs from it.

3.1 Surrounding Dialogue Generation Problem

Before diving into related work to this master thesis, we describe the novel *Surrounding Dialogue Generation Problem* (SDGP), which is derived from the goal of this thesis:

Goal Develop an efficient data-driven architecture which generates, from a dialogue snippet including topic-specific terminology, surrounding topic-specific dialogue suitable as communication exercises for humans after adjustments made by an expert.

The SDGP is defined as follows. Given a dialogue snippet $(u_n, ..., u_m)$ (between two speakers s_1 and s_2), its topic t, and a length l, output an extended dialogue $(u_{n-i}, ..., u_{m+j})$ that is on topic t, has l number of turns, and contains the dialogue snippet $(u_n, ..., u_m)$. It can be divided into the subtasks: find the best placement of the dialogue snippet in the resulting extended dialogue and generate preceding and following utterances to the dialogue snippet (respectively named backward and forward utterance generation). The SDGP is a hard problem, given the current state-of-the-art in NLP. This has implications for both how it is being solved, as discussed in Chapter 4, and how the input and output are treated.

We are interested in machine learning methods to solve the SDGP. In other words, dialogue datasets are the basis for the generation of surrounding dialogues from dialogue snippets.

3.2 Response Generation Systems

A subtask of the SDGP is the response generation task: given a context (usually a dialogue history), generate a response utterance. Many data-driven dialogue systems have been developed to solve the response generation task (e.g., [75, 81, 83, 71, 31]). However, to the best of our knowledge, the generation of preceding utterances has not been researched, an essential subtask of the SDGP. We hypothesize that:

Hypothesis 1 Preceding utterances to a dialogue snippet can be generated using transfer learning techniques similar to those used in response generation systems for following utterances.

Thus, we look into SOTA dialogue systems for response generation, in particular, from the Second Conversational Intelligence Challenge (ConvAI2) [19]. The main task of ConvAI2 was to develop an open-domain, engaging chatbot with a consistent personality prompted at the beginning of the chat. For the competition, a personality is defined through 4-6 short sentences regarding their interests, preferences, and lives. The Persona-Chat dataset [82] (see Table 3.1), which was distributed at the start of the competition, also reflects this [19].

<table>
<tr><th>Persona 1</th><th>Persona 2</th></tr>
<tr><td>I like to ski
My wife does not like me anymore
I have went to Mexico 4 times this year
I hate Mexican food
I like to eat cheetos</td><td>I am an artist
I have four children
I recently got a cat
I enjoy walking for exercise
I love watching Game of Thrones</td></tr>
<tr><td colspan="2">Person 1: Hi
Person 2: Hello ! How are you today ?
Person 1: I am good thank you , how are you?
Person 2: Great, thanks ! My children and I were just about to watch Game of Thrones.
Person 1: Nice ! How old are your children?
Person 2: I have four that range in age from 10 to 21. You?
Person 1: I do not have children at the moment.
Person 2: That just means you get to keep all the popcorn for yourself.
Person 1: And Cheetos at the moment!
Person 2: Good choice. Do you watch Game of Thrones?
Person 1: No, I do not have much time for TV.
Person 2: I usually spend my time painting: but, I love the show.</td></tr>
</table>

Table 3.1: Example dialogue from Persona-Chat. The persona of Person 1 (top left) and Person 2 (top right) is given at the beginning of the chat. This affects their utterances and the topics explored in the dialogue (on the bottom).

ConvAI2 used the automatic metrics used were perplexity, F1 and hits@1/20 to score the competitors, and only the top 7 competitors of the automatic metrics track were evaluated by crowd workers through live interaction with their models [19]. The crowd workers evaluated the chatbots on how much they enjoyed talking to the model and had to verify which persona the model was using given the choice between the correct persona and a random one [19]. Wolf et al. [75] won the automatic metrics track and placed second in the human evaluation track with their proposed TransferTransfo approach [19].

The TransferTransfo approach is to fine-tune GPT-2 [62] on the Persona-Chat dataset with the input representation shown in Figure 3.1 and Figure 3.2. The dialogue state embeddings and special tokens work together to clarify which utterance belongs to which speaker, and what part of the input is the personality of the agent.

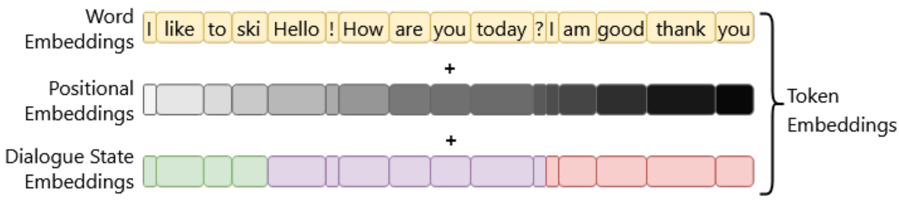


Figure 3.1: The input representation of TransferTransfo, where each token embedding is the sum of a word embedding (top), positional embedding (middle), and dialogue state embedding (bottom). The dialogue state embedding is split into three states: the persona of Person 2 (green), and the dialogue with alternating utterances from Person 1 (purple) and Person 2 (red).

Source: Wolf et al. [75]

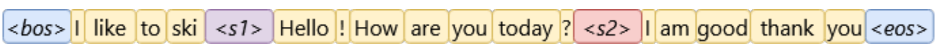


Figure 3.2: The input sequence in the TransferTransfo approach uses beginning-of-sequence (*bos*), end-of-sequence (*eos*), and speaker tokens ($s1$ and $s2$) to mark the different segments of the sequence.

Fine-tuning is done by optimizing over the loss functions: next-utterance classification loss and language modelling loss. *Next-utterance classification loss* is derived from next-sentence classification loss [18] and consists of training a classifier to distinguish a correct next-utterance appended to the input sequence from a set of randomly sampled distractors [75]. The hidden state of the last token (the end-of-sequence token) is passed through a linear layer to get a score, which is applied a cross entropy loss (see Equation 2.4) to correctly classify an answer among distractors. *Language modelling loss* is the commonly used cross entropy loss [75] (see Equation 2.4). To generate utterances in natural language, beam search with sampling [76] is used. The final beams of the beam search is ranked through a scalar combination of the length-normalized utterance probability and the next-utterance classification score [75].

In the light of **Hypothesis 1**, what makes the TransferTransfo approach interesting for this thesis, other than its SOTA performance, is its next-utterance classification loss. The next-utterance classification loss can be transformed into previous-utterance classification loss by trying to classify the correct previous utterance instead of the correct next-utterance. The input sequence can have the utterances in reverse, from the most recent utterance to the oldest. Additionally, its personality consistency property can be used for other aspects of the dialogue (e.g., its topic). This will be further discussed in Chapter 4.

3.3 Human Evaluations for Dialogue Generation

LaMDA [71] (Language Model for Dialogue Applications) is a SOTA large-scale pre-trained language model trained for 57.7 days on a text corpus consisting of 2.97B documents, 1.12B dialogues, and 13.39B dialogue utterances. Like the GPT architectures, it is a decoder-only architecture [71, 61, 62, 9]. What makes LaMDA interesting is that it is fine-tuned on dialogue data with human evaluations of each response in the dataset. The metrics used are:

1. **Sensibleness.** How well the response makes sense in the context of the dialogue (e.g., if the model mentioned that it does not like dogs previously in the dialogue, "I love dogs" is not a sensible response) [3].

2. **Specificity.** How specific the response is (e.g., "Me too" is not specific, but "I love Eurovision songs too" is) [3].

3. **Interestingness.** How interesting the response is (e.g., "Playing the piano is difficult" is not interesting, but "Playing the piano is difficult, since there are so many keys to keep track of, rhythms, and key combinations, but after enough time, you just know by feeling" is interesting [71].

4. **Safety.** How safe the response, such that it does not harm anyone nor have an unfair bias [71]. This metric is defined as several, specific safety objectives.

5. **Groundedness.** The percentage of utterances that carry information on the external world that can be supported by external sources [71].

6. **Informativeness.** The percentage of responses that carry information on the external world that can be supported by the other utterances in the dialogue [71].

7. **Citation accuracy.** The percentage of responses that cite the URLs of their sources as a share of all responses with explicit claims about the external world, excluding claims with well-known facts (such as "horses have four legs") [71].

8. **Helpfulness.** How helpful the response is in the context of the user's needs, given that they contain correct information (based on the user's independent research) [71].

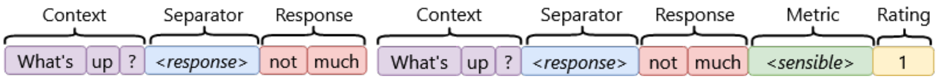
9. **Role consistency.** How consistent the model's response is to its role external to the dialogue (e.g., an assistant that is aware that it is not human) [71].

The human evaluations were collected using crowd workers. Crowd workers interact with a LaMDA instance and mark the response as sensible, specific, and/or interesting given a context (with binary labels), resulting in 6.4K dialogues (61K turns) with annotated LaMDA-responses. Human-annotated dialogues are also collected similarly for safety (8K dialogues) with binary labels for each of the safety objectives. Finally, 4K dialogues are annotated on groundedness by having crowd workers write queries to an information retrieval system and modify the model responses.

LaMDA is fine-tuned on two tasks: the generative task and the discriminative task. The generative task is to generate a response given a context (a dialogue history). The discriminative task is to predict the rank of a response on a metric given a context. The input representations and where the losses are applied in the sequences are seen in Figure 3.3.

LaMDA is fine-tuned to predict the sensibleness, specificity, interestingness, and safety of generated responses (using the dialogue data with binary labels on each of the metrics), though the safety ratings are aggregated into a single metric. When the fine-tuned LaMDA is used to respond to, e.g., a user's utterances it generates candidate responses, predicts their ranks, filters away believed unsafe responses (i.e., when the safety prediction rank falls below a threshold) and selects the top-ranked candidate as the next response. The best weighted sum ranking strategy found is that each candidate response's rank is:

$$rank(r) = 3 * P(sensible) + P(specific) + P(interesting), \tag{3.1}$$



(a) Generative task (b) Discriminative task

Figure 3.3: For fine-tuning LaMDA, the input representation for the generative task (to the left) is a sequence of tokens expressed as context, separator, and response. Losses are only applied to the response portion, such that the model learns to generate the responses only given the context. While in the discriminative task (to the right), the input sequence includes the metric name and rating. Losses are only applied to the rating following the metric name, so the model learns to predict the rating of the response for the given metric and context.

where r is the candidate response and $P(metric)$ is the predicted rank of the response on the metric given the context [71]. This fine-tuning approach is expensive, as it uses data with human evaluations, but achieves great results with the ability to both generate and rank its generated responses.

Additionally, LaMDA is fine-tuned to learn to call an external information retrieval system to boost the groundedness of its responses, using the dialogue data with groundedness annotations. Since we do not focus on the groundedness of the generated utterances in this thesis, we will not go into detail on this aspect. Essentially, the fine-tuned LaMDA model can interact with a toolset consisting of an information retrieval system, a calculator, and a translator through strings. For instance, the translator takes “hello in French” and outputs “Bonjour” in a list. The model will augment its response accordingly, to the information it receives from the toolset. If the input cannot be parsed by any tool, it will return an empty list and does not contribute to the final response.

The fine-tuned LaMDA was evaluated by crowd workers on sensibleness, specificity, interestingness, informativeness, safety, and groundedness. The crowd workers labelled (with binary labels) LaMDA-generated responses given the context of these metrics. The results showed that LaMDA can achieve near-human performance on sensibleness, specificity and interestingness [71]. While fine-tuning of safety showed increased performance, LaMDA's gap to human performance is still significant [71]. We believe this may be due to the complexity and limitations of the safety metrics, making the collected evaluation data from the crowd workers prone to errors in the form of different interpretations and strictness. While fine-tuning of LaMDA to generate calls to an external information retrieval system showed increased performance, its gap to human performance is still significant [71]. Thus, as the informativeness metric is similar to groundedness, we hypothesize that:

Hypothesis 2 The evaluation metrics used in LaMDA can be used to evaluate dialogues by transforming them from utterance-level to dialogue-level and result in high-quality evaluations for sensibleness, specificity, and interestingness, but may be lacking in groundedness and informativeness.

This is because, like the safety metric, the groundedness and informativeness metrics may be more difficult to understand and, thus, give evaluations on, resulting in unreliable evaluation data (and more noise in the training set). Also, when evaluating dialogues in the context of surrounding dialogue generation, the final dialogue is more interesting, rather than the context between each of the individual utterances. Even though we do not focus on optimizing the groundedness and informativeness of generated dialogues, it is still interesting to measure across these metrics to see if there is room for improvement.

3.4 Adapter-based Tuning

The two previous works used fine-tuning to adapt a large-scale language model to the specific task of response generation and the prediction of the rank of a response. A task-oriented dialogue system incorporating continuous learning through adapter-based tuning is AdapterCL (Adapter Continuous Learning) [47]. In this section, we describe the AdapterCL method, continuous learning in task-oriented dialogue, and the benefits of adapter-based tuning.

Continuous learning is to learn a set of tasks sequentially without catastrophically forgetting previously learned ones [48, 72]. Madotto et al. [47] defined continuous learning in task-oriented dialogue systems as learning a sequence of domains sequentially (see Figure 3.4). We are interested in this since the SDGP consists of multiple tasks, forward utterance generation and backward utterance generation.

In AdapterCL, GPT-2 is adapter-based tuned with an adapter for each domain (see Figure 2.6) and a perplexity-based classifier is used to select which adapter to activate and use at testing time. In other words, to learn a new domain, a new adapter is placed in-between the layers of GPT-2 and trained over the domain-specific dataset. This is done with Residual Adapters [32] over the commonly used language modelling loss function (see Equation 2.4). Given a set of adapters parameterized by $\mu_0, ..., \mu_N$, each trained respectively on the domain-specific datasets $D_0, ..., D_N$, and an input sample X, the classifier computes:

$$\alpha_t = \mathrm{PPL}_{\mu_t}(X), \tag{3.2}$$

for $t = 1, ..., N$, where $\mathrm{PPL}_{\mu_t}(X)$ is the perplexity of the model μ_t on the input sample X (see Equation 2.7). Here, α_t represents the confidence of adapter t

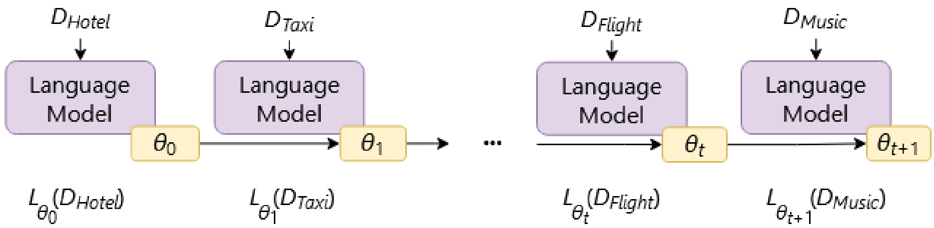


Figure 3.4: In continuous learning for task-oriented dialogue systems, the model is trained on one dataset at the time (each for a specific domain). Here, the model is trained on data from the hotel domain D_{Hotel}, then the taxi domain, and so on. The parameters of the model θ are updated sequentially based on the loss function L_θ.

Source: Madotto et al. [47]

being appropriate for the input X. The adapter t with the smallest α_t is selected to perform the task.

AdapterCL outperformed most baselines on intent recognition (knowing what the user wants from the task-oriented dialogue system) and on slot error rate (for the natural language generation task) over time [47]. Most baselines struggled to keep up to AdapterCL the more domains and tasks they learned.

AdapterCL can be seen as a mixture of experts, with an expert (or rather adapter) for each domain. A mixture of experts divides the problem space into smaller, more manageable subsets to be mastered by their own expert. This is a learning methodology that was proposed in the 1990s [35]. An interesting discussion is how coupled the experts should be given a problem space. While Jacobs and Jordan [34] and Hampshire and Waibel [27] suggested cooperative experts, competitive experts were proposed by Jacobs et al. [35] for less inference. In AdapterCL, the experts are competitive, since the experts do not work together to generate an output. The mixture of experts application in AdapterCL makes the architecture more modular. AdapterCL uses this for continuous learning. Adapters for new domains to be added to the architecture later, sequentially, given that the classifier can handle it. While this is also possible with fine-tuning, a model for each domain, it is less efficient [32]. Adapter-based tuning also outperforms fine-tuning on low-resource tasks and is more robust to overfitting and less sensitive to changes in learning rates [28]. Thus, we hypothesize that:

Hypothesis 3 An adapter can be assigned to each subtask of the SDGP to develop an efficient language model for surrounding dialogue generation.

Chapter 4

Proposed Architecture

Our novel BFD Generator (Backward-Forward Dialogue Generator) architecture is presented in this chapter. It is an SDGP architecture that generates an extended dialogue from a dialogue snippet and additional information, as seen in Figure 4.1, using forward and backward generation.

In Section 4.1, the datasets considered and used in the architecture are presented. In Section 4.2, the performance phase for our architecture is described, and in Section 4.3 the learning phase. The rest of our architecture's components are described in Section 4.2.1 and Section 4.2.2. In Section 4.4, we discuss design choices for our architecture. Finally, we discuss how our architecture differs from related work in Section 4.5.

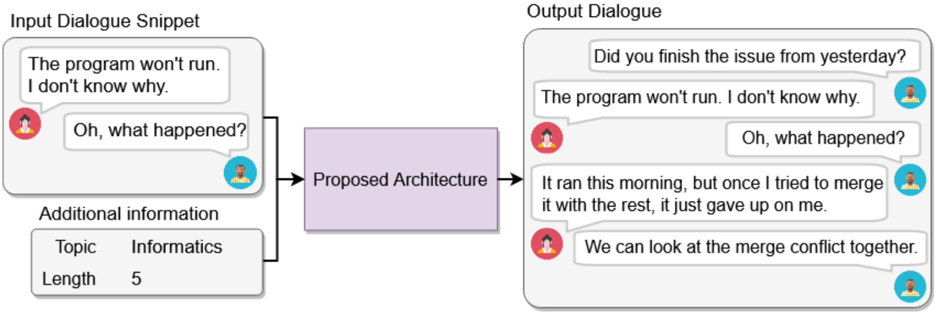


Figure 4.1: The proposed architecture (BFD Generator) takes a dialogue snippet and additional information (topic and length) as input to generate a more complete dialogue as output.

4.1 Datasets

In ML, the data is the foundation for learning. Thus, finding a suitable dataset is one of the most important tasks. Not only does the data need to be plentiful, it needs to contain texts of the appropriate style and characteristics. This section presents the dataset containing content from the company we collaborate with and the dataset we use for training our architecture, Wizard of Wikipedia dataset [20].

Dialogue Exercise Corpus The company has created around 50 dialogues for language learning in their app. We refer to it as the Dialogue Exercise Corpus. Each dialogue can branch in different directions like a graph, simulating a close to real-life dialogue experience where users' actions are responded with appropriate events in the app. Some branches of dialogues may also reunite later. The dialogues concern different domains and CEFR levels, and are tailored to their customers. Each dialogue is connected to a scenario, meaning they all have a goal in mind. A dialogue scenario contains pre-determined exchanges between the user and another speaker, as well as a brief (describing the beginning of the scenario) and debrief (describing how the scenario continued and feedback on how well the user performed) [12].

While the Dialogue Exercise Corpus contains the type of dialogues they wish to generate, it is data-sparse. Training a model on it has a high risk of overfitting. Thus, it is only used to study the values and writing style of the company.

Training Corpus The conversational and open-domain dataset with topic-labelling Wizard of Wikipedia [20] is chosen for adapting the language model. We denote the dataset D, and its data points d. Each data point d consists of a topic t (a noun), a dialogue consisting of multiple utterances $(u_0, ..., u_N)$, and two speakers $(s1, s2)$. A sample of the dataset is seen in Figure 4.2.

The dialogues in Wizard of Wikipedia were collected by having crowd-workers engage in 1-to-1 dialogues. One of the crowdworkers picks a topic to chat about. One crowdworker plays the role as an apprentice, who is eager to learn more about the topic. The other crowdworker plays the role as a wizard, who knows that their chatting partner is eager to discuss about the topic. For each turn, the wizard is shown relevant knowledge (extracted from Wikipedia) and can choose a relevant sentence to construct a response [20]. Otherwise the wizard just constructs a sentence on their own. Thus, at least one of the speakers are knowledgeable about the topic, which can simulate jargon or explanation of jargon in different scenarios (e.g. at workplaces). The dataset is also available for commercial use, which is a requirement from the company we collaborate with.

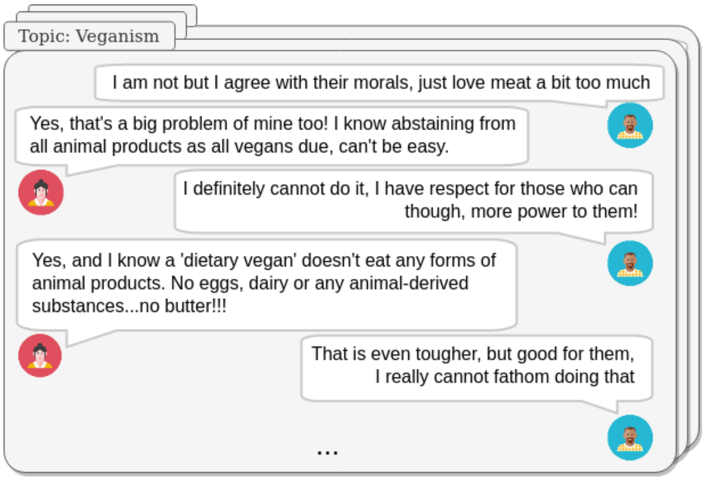


Figure 4.2: Each data sample in WIZARD OF WIKIPEDIA contains a topic (e.g. Veganism) and exchanges between the two speakers.

4.2 BFD Generator: Performance Phase

This section describes the performance phase of the BFD Generator by going through Algorithm 1. The BFD Generator extends a dialogue snippet $\mathcal{S}$ using forward and backward utterance generation. The BFD Generator consists of a pre-trained causal language model M that can switch between using two adapters A_B and A_F (see line 1-4 in Algorithm 1), a decoding method (see line 10 in Algorithm 1) and a selection module (see line 29 in Algorithm 1).

The BFD Generator input consists of a dialogue snippet $\mathcal{S}$, topic t, and dialogue length l in turns (see line 5 in Algorithm 1). The dialogue snippet $\mathcal{S}$ consists of m utterances u_j, as seen in Equation 4.1. The utterances alternate between the two speakers s_1 and s_2.

$$\mathcal{S} = \begin{cases} u_n & \text{if } m = n \\ (u_n, ..., u_m) & \text{if } m > n \end{cases} \tag{4.1}$$

The main function of the BFD Generator is BFD-GENERATOR($\mathcal{S}, t, l$) on line 5-9. In each iteration of the while-loop, candidate utterances are generated from the snippet $\mathcal{S}$ and topic t (see line 7 in Algorithm 1) and the best candidate is added to the snippet (see line 8 in Algorithm 1). This is repeated until the length of the snippet $\mathcal{S}$ is equal to the input length l. An example of Algorithm 1 running is shown in Figure 4.3.

The candidates are generated in the CANDIDATE-GENERATOR($\mathcal{S}, t$) function using the causal language model M, the adapters A_B and A_F, and the decoding method (defined in line 20-28). First, the adapter for forward utterance generation is activated in the language model M. Then k candidates for following utterances to the snippet $\mathcal{S}$ is generated and added to the array *candidates* (see line 13-14 in Algorithm 1). The process is repeated for backward utterance generation, except we reverse the order of the utterances in the snippet (see line 15 in Algorithm 1) first. Finally, the preceding and following candidate utterances are returned.

The utterances are generated token-by-token in DECODING-METHOD($\mathcal{S}, t, M$) (see line 20-28 in Algorithm 1). We do this by sending in the snippet $\mathcal{S}$ and *topic* to the adapter-based tuned language model. The language model then outputs the token probability distribution, which is decoded using top-p sampling with temperature θ. The chosen token is added to the utterance. This sequence is repeated until the chosen token *token* is a speaker token ($s1$,$s2$), which marks the end of the utterance. The utterance is finally returned.

After the candidates have been generated, we score each of the candidate utterances and add the best candidate in the SELECTION-MODULE($\mathcal{S}$, t) function (see line 29-44 in Algorithm 1). A more in-depth explanation of the scoring measures (see line 36-38 in Algorithm 1) is given in Equation 4.5. The highest scoring candidate is added to the snippet $\mathcal{S}$ (see line 43 in Algorithm 1) in its appropriate place (see Equation 4.2). Finally, the snippet extended by an utterance $\mathcal{S}$ is returned.

$$\mathcal{S} = \begin{cases} (u^i_{n-1}, u_n, ..., u_m) & \text{if preceding utterance } u^i_{n-1} \text{ is selected} \\ (u_n, ..., u_m, u^i_{m+1}) & \text{if following utterance } u^i_{m+1} \text{ is selected} \end{cases} \tag{4.2}$$

Algorithm 1 BFD-Generator (Performance Phase)

global variables
  M, causal language model.
  A_B, adapter for backward utterance generation.
  A_F, adapter for forward utterance generation.

---

function BFD-GENERATOR($\mathcal{S}, t, l$) ▷ Extend a dialogue snippet $\mathcal{S}$ on topic t to a dialogue of length l (in turns) through surrounding dialogue generation.
  while $\mathcal{S}$.Length $< l$ **do**
    $candidates \leftarrow$ CANDIDATE-GENERATOR($\mathcal{S}$,t)
    $\mathcal{S} \leftarrow$ SELECTION-MODULE($candidates, \mathcal{S}$)
  return $\mathcal{S}$

---

function CANDIDATE-GENERATOR($\mathcal{S}, t$) ▷ Generate k candidate utterances for each direction.
  let k be a constant number in $\mathbb{N}$.
  $M \leftarrow$ SET-ACTIVE-ADAPTER(A_F) ▷ Set A_F as the active adapter in M.
  for $j \leftarrow 1,\, k$ **do**
    $candidates \leftarrow candidates \cup$ DECODING-METHOD($\mathcal{S}$,t,M)
  $\mathcal{S}_r \leftarrow$ REVERSE($\mathcal{S}$) ▷ Order the utterances in $\mathcal{S}$ from most recent to oldest.
  $M \leftarrow$ SET-ACTIVE-ADAPTER(A_B) ▷ Set A_B as the active adapter in M.
  for $j \leftarrow 1,\, k$ **do**
    $candidates \leftarrow candidates \cup$ DECODING-METHOD($\mathcal{S}_r$,t,M)
  return $candidates$

---

function DECODING-METHOD($\mathcal{S}, t, M$) ▷ Generate an utterance. See Section 4.2.1.
  let p and θ be constant numbers between 0 and 1.
  $u \leftarrow \emptyset$
  repeat
    $P \leftarrow$GET-TOKEN-PROBABILITY-DISTRIBUTION(M,$\mathcal{S}$,t)
    $token \leftarrow$TOP-P-SAMPLING(P,p,θ) ▷ With temperature θ.
    $u \leftarrow u \cup token$
  until $token$ **is** $s1$ or $s2$
  return u

---

function SELECTION-MODULE($candidates$, $\mathcal{S}$) ▷ Score the candidates and add the best to dialogue snippet $\mathcal{S}$. See Section 4.2.2.
  let $g_{\text{self}}, g_{\text{response}}, k_1, k_2, k_3, k_4, k_5$ be constant positive numbers.
  $score_{best} \leftarrow -\infty$
  $u_{best} \leftarrow \emptyset$
  for each $u_c \in candidates$ **do**
    let U^{self} be the utterances in $\mathcal{S}$ by the speaker of the candidate u_c
    let U^{response} be the utterances in $\mathcal{S}$ not by the speaker of the candidate u_c
    $n_c \leftarrow$ PROPER-NOUN-COUNT(u_c)
    $d_{\text{self}} \leftarrow \sum_{i=0}^{|U^{\text{self}}|}(g_{\text{self}} - sim(u_c, u_i) * k_4^i)$ ▷ Use cosine similarity between two utterances $sim(u_c, u_i)$
    $d_{\text{response}} \leftarrow \sum_{i=0}^{|U^{\text{response}}|}(g_{\text{response}} - sim(u_c, u_i) * k_5^i)$
    $score \leftarrow k_1 * n_c - k_2 * d_{\text{self}} - k_3 * d_{\text{response}}$
    if $score > score_{best}$ **then**
      $score_{best} \leftarrow score$
      $u_{best} \leftarrow u_c$
  $\mathcal{S} \leftarrow \mathcal{S} \cup u_{best}$ ▷ Add the best candidate to the dialogue snippet $\mathcal{S}$ at the appropriate place.
  return $\mathcal{S}$

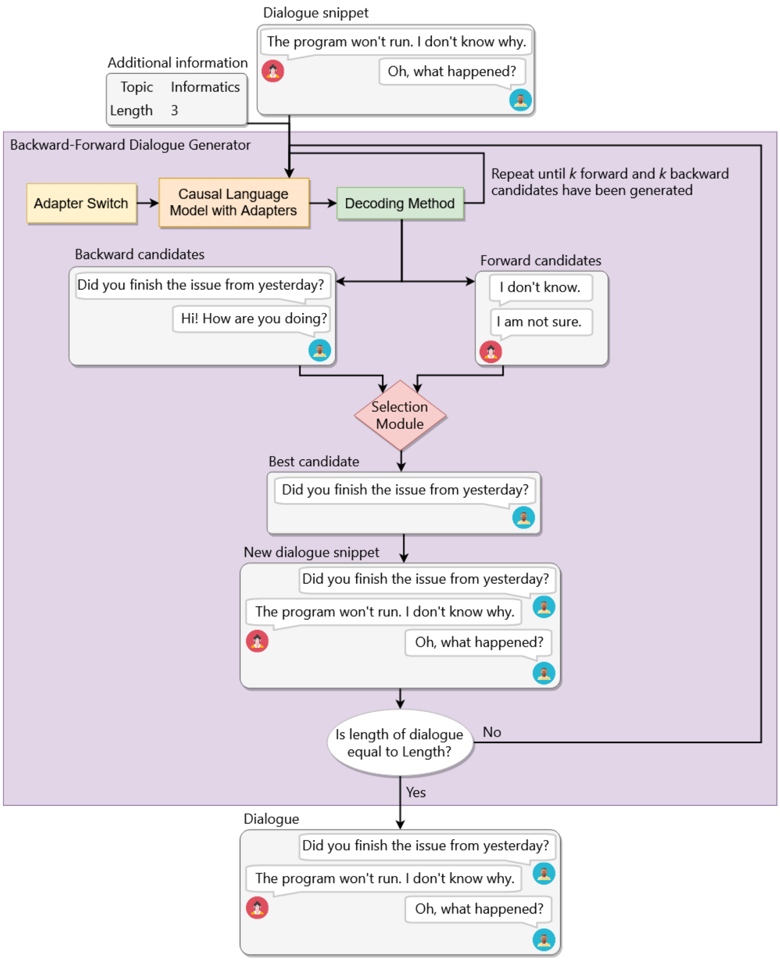


Figure 4.3: Performance phase pipeline of the BFD Generator. The input (see Figure 4.1) is a dialogue snippet and additional information (i.e., topic, length). The snippet and topic are sent to the causal language model to generate $2k$ candidate utterances in natural language using the decoding method and the adapters for forward and backward utterance generation. The Adapter Switch is used to activate the appropriate adapter before generating an utterance. The candidates are then scored in the Selection Module. The best candidate is added to the snippet. These steps are repeated until the length of the snippet is equal to the input length. Finally, the extended dialogue is output (see Figure 4.1).

4.2.1 Decoding Method

Dialogue generation is an open-ended task, where each dialogue snippet can have many good preceding and following utterances. That is one difficulty in generating dialogues. It is also difficult to judge if an utterance is suitable for a dialogue snippet in the middle of constructing the utterance. To tackle these difficulties, we generate a pool of candidate utterances of which we can compare to the dialogue snippet and against each other more easily. To keep the candidates diverse while staying on-topic, we use top-p sampling with temperature θ as the decoding method (see line 25 in Algorithm 1). This is because it is more prone to generating more natural, engaging and interesting text, while reducing the chance of trailing off the topic [30, 49]. In addition, we make sure candidates (for each direction) are different from each other.

In the adapter-based tuning process of the causal language model, we do not swap between which speaker starts a dialogue. This means the model's performance may rely on appropriate assignment of the speakers. To handle the assignment of speakers, we generate half of the candidates with one speaker first and the rest with the other speaker first. This means that in one of the two scenarios, the assignment of the speaker will be appropriate.

4.2.2 Selection Module

The Selection Module (SM) takes the candidates (utterances), scores them according to a weighted sum equation, and selects the highest scoring candidate to be added to the dialogue depending on their relevance to the dialogue snippet. Due to the independence of the adapters A_B and A_F, only one of the candidates (instead of one from each adapter) is added to the dialogue (see line 43 in Algorithm 1). This is to avoid discrepancies in the dialogue since the dialogue snippet changes when a candidate is added, but the candidates themselves were generated before this change. However, once the forward and backward candidates are far enough from each other in the dialogue, this is no longer a problem due to the Markov assumption (see Section 2.3). For simplicity, we will only add one candidate at a time in each iteration of the BFD Generator. The score for each candidate is given by a weighted sum of its proper noun count and utterance similarities, both of which are defined below.

Proper Noun Count Proper noun count is a simple metric counting the number of proper nouns. Consecutive nouns, of which at least one is a proper noun, like "Google Store" and "Jack Sparrow" count as one proper noun each.

Utterance Similarity Utterance similarity is split into self similarity d_{self} and response similarity $d_{response}$. Self similarity is the utterance's similarity to the speaker's utterances in the dialogue snippet. While response similarity is the utterance's similarity to the utterances by the other speaker. We distinguish these similarities as they may have different impacts, which we can research.

Let U^{self} be the set of all utterances u_i in the dialogue snippet which origin from the speaker of the candidate u_c. Let $sim(u_c, u_i)$ denote the cosine similarity between the utterances u_c and u_i which are encoded with the Universal Sentence Encoder [11]. The self similarity d_{self} between the candidate u_c and the dialogue snippet with the self utterances u_i for $i = 1, ..., |U^{\text{self}}|$ is given in Equation 4.3.

$$d_{\text{self}}(u_c, U^{\text{self}}) = \sum_{i=0}^{|U^{\text{self}}|} (g_{\text{self}} - sim(u_c, u_i) * k_4^i), \tag{4.3}$$

where $g_{\text{self}}, k_4 \in (0, 1)$ are coefficients. Let U^{response} be the set of all utterances in the dialogue snippet that are from the other speaker. The response similarity between the candidate u_c and the dialogue snippet with the utterances u_i for $i = 1, ..., |U^{\text{response}}|$ is given in Equation 4.4.

$$d_{\text{response}}(u_c, U^{\text{response}}) = \sum_{i=0}^{|U^{\text{response}}|} (g_{\text{response}} - sim(u_c, u_i) * k_5^i), \tag{4.4}$$

where $g_{\text{response}}, k_5 \in (0, 1)$ are coefficients. g_{self} and g_{response} denote the gold (ideal) similarity score between the candidate and each utterances, for self and response utterances respectively. k_4 and k_5 are coefficients to reduce the influence of utterances that are further away from the candidate. This is because if the candidate is closely related to the preceding and/or following utterance, it can result in a more natural flow in the dialogue. Thus, the closer an utterance is to the candidate, the more impact to the similarity score we give it.

Candidate Score Let n_c be the proper noun count of the candidate u_c. The score for a candidate u_c with the dialogue snippet $\mathcal{S} = U^{\text{self}} \cup U^{\text{response}}$ is given by the weighted sum $s(u_c, \mathcal{S})$, seen in Equation 4.5:

$$s(u_c, \mathcal{S}) = k_1 * n_c - k_2 * d_{\text{self}}(u_c, U^{\text{self}}) - k_3 * d_{\text{response}}(u_c, U^{\text{response}}) \tag{4.5}$$

where k_j for $j = 1, 2, 3$ are positive coefficients. Due to the Markov assumption (see Section 2.3), we can limit the number of utterances to consider for practicality and run-time. Thus, we do not consider the utterances that are the furthest away (limited by some parameter) from the candidate for calculating the self and response similarity. However, issues with this approach are long-range dependencies (e.g., distant, but relevant utterances) and long-range use of proper nouns.

4.3 BFD Generator: Learning Phase

This section describes the learning phase of the BFD Generator, i.e., our transfer learning approach for the causal language model, by going through Algorithm 2. We adapt the pre-trained causal language model GPT-2 using adapter-based tuning [6, 32]. We train an adapter for forward utterance generation and another adapter for backward utterance generation.

The main function of the learning phase of the BFD Generator is TRAIN-ADAPTERS(D) (see line 5-15 in Algorithm 2). First, we set up the pre-trained language model M and pre-trained tokenizer T (see line 8 in Algorithm 2). Afterward, the dataset D is pre-processed (see line 9 in Algorithm 2). The pre-processed data D transformed to a training set *inputs* for forward utterance generation (see line 10 in Algorithm 2). The training set *inputs* is used to train the adapter for forward utterance generation A_F using the language model M as the base model (see line 11 in Algorithm 2). When the adapter has been trained, it is saved (see line 12 in Algorithm 2) so it can be loaded for the performance phase later. To train the adapter for backward utterance generation, we follow the same steps (see line 13-15 in Algorithm 2). However, the training set *inputs* is set up for backward utterance generation instead (see line 13 in Algorithm 2).

To set up the language model M and tokenizer T, we use *checkpoint* to retrieve the pre-trained models from an external source, e.g., a library (see line 17-18 in Algorithm 2). The vocabulary of the language model M and tokenizer T is extended (see line 19 in Algorithm 2) to include our special tokens, the beginning-of-sequence token *bos*, end-of-sequence token *tos*, and the speaker tokens $s1$ and $s2$. Otherwise, they may misunderstand the special tokens. Before returning the language model M and tokenizer T, a language modelling head and a classification head are added to the language model M for multi-task learning (see Section 4.3.2).

We perform simple pre-processing of the dataset D (see line 23-26 in Algorithm 2). This is described in Section 5.1.

In the SET-UP-INPUTS($D, T, isForward$) function (see line 27-40 in Algorithm 2), the training set *inputs* is set up. For each of the data points d in D, we create training data points. Each of the training data points in *inputs* consists of a dialogue snippet $\mathcal{S}$, a topic t, a target utterance $targetUtterance$, and distractors $distractors$. For each dialogue containing N utterances, we create $N-1$ snippets $\mathcal{S}$ (see line 36 in Algorithm 2) with a target utterance $targetUtterance$ each (see line 35 in Algorithm 2). The target utterance is the following utterance to the snippet $\mathcal{S}$ if the boolean $isForward$ is true (i.e., when we set up a training set for forward utterance generation). If $isForward$ is false, then the target utterance is the preceding utterance to the snippet $\mathcal{S}$ and the snippet $\mathcal{S}$ is ordered from the most recent to oldest utterance. The final step of setting up the training set is to represent *inputs* as input sequences (see line 39 and line 41-47 in Algorithm 2). The input representation is explained in Section 4.3.1. The padded training set (see line 46 in Algorithm 2) with our input representation is returned. Padding is a technique to make the inputs equal in dimensions, otherwise the neural network cannot handle the inputs.

Once the training set is ready, we train the adapter A_F for forward utterance generation (see line 11 in Algorithm 2). First, we activate the adapter A_F in the language model M. Then, we freeze the layers of the base language model M. Finally, we train the language model M, with the activate adapter A_F on the training set *inputs*. The model M with the trained active adapter A_F is returned. To train the adapter A_B for backward utterance generation (see line 14 in Algorithm 2), the same steps are followed.

4.3.1 Input Representation

Since GPT-2 is a decoder-only architecture, all inputs for training the language model are expressed as sequences of tokens. We use special tokens (see line 4 Algorithm 2) to separate the segments of the input sequences (see line 42 in Algorithm 2). The structure of the input representation is illustrated in Figure 4.4 with the special tokens: *bos*, s_1, s_2, and *eos*. The *bos*- and *eos*-tokens are necessary to mark the beginning and end of a sequence, respectively. The speaker tokens s_1 and s_2 segment the sequence by indicating which utterances belong to which speaker. This allows the model to learn consistencies between the speakers. This is kept consistent for each dialogue (and not only for each input sequence). For instance, one of the speakers may be knowledgeable about the topic, while the other wants to learn about it. It is important to keep the speakers distinct to make it possible to learn this distinction. The speaker tokens also separate the topic from the rest of the dialogue.

Word, Positional and Segment Embeddings We use word, positional and segment embeddings in the input representation (see Figure 4.5 and line 43-45 in Algorithm 2) [75]. Each token gets a word embedding from the pre-trained tokenizer T to represent the meaning of the word. Since the language model is a symmetrical dot-product, we also add positional information of each token. This way, the position of words are represented. In addition to the speaker-tokens in the input sequence (see Figure 4.4), we have segment embeddings consisting of the speaker-tokens only. This is to strengthen the indication of which segment each word belongs to [75].

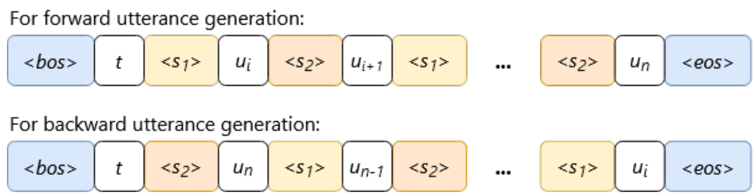


Figure 4.4: The structure of the input sequence for tuning of the forward adapter (top row) and the backward adapter (bottom row) using special tokens as delimiters. Each of the tuning sequences sent into the causal language model are wrapped by a *bos*-token (beginning-of-sequence) and an *eos*-token (end-of-sequence). Following the *bos*-token, we have the topic t and the dialogue snippet consisting of the utterances $(u_n, ..., u_m)$ which alternate between the two speakers, indicated with the speaker-tokens s_1 and s_2. Notice how the order of the utterances are reversed between the forward and backward input sequences.

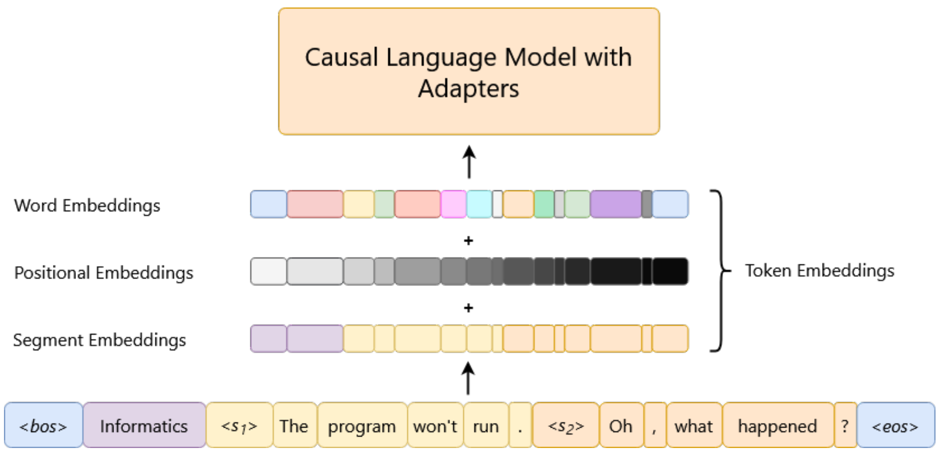


Figure 4.5: The input representation of the causal language model for adapter-based tuning is a token embedding consisting of word, positional and segment embeddings.

Source: Wolf et al. [75]

4.3.2 Next-Sentence Prediction and Language Modelling

We use multi-task learning to train the adapters over the two tasks: language modelling and next-sentence prediction. For language modelling, the adapters are optimized over the commonly used cross-entropy loss (see Equation 2.4). We divide the next-sentence prediction task into: the preceding-utterance prediction task and the following-utterance prediction task (also known as next-utterance prediction task from Section 3.2). Their only difference is that in preceding-utterance prediction, the goal is to correctly classify the preceding utterance to a dialogue among distractors (see line 37 in Algorithm 2). Distractors are utterances randomly sampled from the dialogue dataset, which are not equal to the correct preceding utterance of the task.

Algorithm 2 Train-Adapters (Learning Phase)

```
global variables
    checkpoint, refers to a pre-trained causal language model and tokenizer.
    s1 and s2, speaker tokens.
    specialTokens, an array of the special tokens s1, s2, bos, eos.

function TRAIN-ADAPTERS(D)    ▷ Adapter-based tune the causal language model for forward and backward utterance generation.
    let A_B and A_F be untrained adapters.
    let checkpoint refer to a pre-trained causal language model and tokenizer.
    M,T ← SET-UP-LANGUAGE-MODEL-AND-TOKENIZER(checkpoint)
    D ← PRE-PROCESS-DATA(D)
    inputs ← SET-UP-INPUTS(D,T, true)
    M ← TRAIN-ADAPTER(M, A_F, inputs)
    SAVE-ADAPTER(M)    ▷ Save the active adapter.
    inputs ← SET-UP-INPUTS(D,T, false)
    M ← TRAIN-ADAPTER(M, A_B, inputs)
    SAVE-ADAPTER(M)

function SET-UP-LANGUAGE-MODEL-AND-TOKENIZER(checkpoint)
    M ← GET-PRE-TRAINED-LANGUAGE-MODEL(checkpoint)
    T ← GET-PRE-TRAINED-TOKENIZER(checkpoint)
    M,T ← ADD-SPECIAL-TOKENS(M,T,specialTokens)
    M ← ADD-LANGUAGE-MODELLING-HEAD(M)    ▷ See Section 4.3.2.
    M ← ADD-CLASSIFICATION-HEAD(M)    ▷ See Section 4.3.2.
    return M, T

function PRE-PROCESS-DATA(D)    ▷ Pre-process the WIZARD OF WIKIPEDIA dataset.
    D ← REMOVE-OUTLIERS(D)    ▷ See Section 5.1.
    D ← REMOVE-REDUNDANT-WHITESPACES(D)
    return D

function SET-UP-INPUTS(D, T, isForward)    ▷ Setup the input representation for training. We assume that isForward is true (see Section 4.3 for explanation of how false is handled).
    let s1 and s2 be the speaker tokens.
    inputs ← ∅
    for each d ∈ D do
        t ← GET-TOPIC(d)
        S ← ∅
        let the utterances in d be ordered from the oldest to the most recent.
        for each utterance ∈ d do    ▷ However, skip the final utterance.
            let targetUtterance be the utterance following utterance in the dialogue d.
            S ← S ∪ utterance
            distractors ← GET-DISTRACTORS(targetUtterance)    ▷ See Section 4.3.2.
            inputs ← inputs ∪ (t, S, targetUtterance, distractors)
    inputs ← ADD-SPECIAL-TOKENS-EMBEDDINGS-AND-PADDINGS(inputs,T)
    return inputs

function ADD-SPECIAL-TOKENS-EMBEDDINGS-AND-PADDINGS(inputs,T)    ▷ See Section 4.3.1.
    inputs ← ADD-SPECIAL-TOKENS-TO-INPUTS(inputs,specialTokens)
    inputs ← ADD-WORD-EMBEDDINGS(inputs,T)    ▷ By using the tokenizer T
    inputs ← ADD-POSITIONAL-EMBEDDINGS(inputs)
    inputs ← ADD-SEGMENT-EMBEDDINGS(inputs, s1, s2)
    inputs ← PAD-DATASET(inputs)
    return inputs

function TRAIN-ADAPTER(M, A, inputs)    ▷ Train adapter A on the dataset inputs.
    M ← SET-ACTIVE-ADAPTER(A)    ▷ Set A as the active adapter in M.
    M ← FREEZE-LAYERS(M)    ▷ Freeze the base language model's layers.
    M ← TRAIN(M,inputs)    ▷ Train the language model M.
    return M
```

4.4 BFD Generator: Design Discussion

In this section, we discuss our design choices. This includes choice of dataset, pre-trained language model, and overall approach.

There are many conversational dialogue datasets available. Most of them are open-domain [41, 82, 20] and have a specific focus. Some focus on personalities, both in writing style and background information like occupation and gender (e.g., PersonaChat [82]). Other datasets focus on emotions (e.g., DailyDialog [41] and EmotionLines [33]). However, while personality and emotions of the speakers can add depth to the dialogues, we focus on the topic of the dialogue. We hypothesize that specialized content can be more consistently generated by having the topic of a dialogue as a condition for generating the surrounding dialogue. In addition, for educational content creation, by keeping the topic consistent throughout the dialogues, people can learn more relevant phrases and responses regarding the topic. Thus, they can increase their language proficiency in that particular topic with more efficiency. Therefore, we use the Wizard of Wikipedia dataset to adapt the causal language model since it has conversational dialogues labelled by topic. In addition, it is available for commercial use, which is a requirement from the company we collaborate with.

A downside with Wizard of Wikipedia is that it mostly contains dialogues on the CEFR level B2, grammatically. This makes it difficult to train a model to generate dialogues with different CEFR levels. Thus, we do not attempt to adapt the causal language model to generate utterances on a specific CEFR level. We expect most of the generated utterances to be on a B2 level.

The causal language model GPT-2 is chosen for the BFD Generator. Its causal characteristic makes it and expert in predicting the next word given a dialogue snippet. GPT-2 has also shown promising results for dialogue generation [75]. We use its strength to write utterances in a forward manner, word by word, and adapt it to understand the dependency between utterances in a forward and backward manner. GPT-2 is free, open-source, and has a manageable amount of parameters, in contrast to its succeeder GPT-3. It has shown promising results as the core of various conversational and task-oriented dialogue systems [75, 31]. However, the modularity of the BFD Generator allows any kind of causal language model to be used. This adds to the sustainability of our architecture, since GPT-2 can be swapped out when causal language models with increased performance are released.

By taking a modular approach, using separate adapters for forward and backward utterance generation, it allows for more explainable ablation testing. We hypothesize that adapter-based tuning requires less resources to train to an adequate level and more space efficient than fine-tuning. Adapter-based tuning also allows for swapping of active experts in the large-scale language model for forward and backward utterance generation.

Adapter-based tuning is a modular approach. It allows for swapping of experts while only using one large-scale language model. This is space efficient. We use this characteristic of adapter-based tuning to swap between the experts for forward and backward utterance generation.

4.5 From Related Work to BFD Generator

In this section, we discuss how the BFD Generator differs from the approaches it is inspired from. In particular, it is compared to the related work from Chapter 3.

TransferTransfo [75] The BFD Generator uses adapter-based tuning, while TransferTransfo uses fine-tuning. TransferTransfo is only for forward utterances, while the BFD Generator is for both forward and backward utterances (where backward utterances are made possible by training on dialogue histories from the most recent to the oldest utterance. The BFD Generator uses topics (noun) instead of personas (4-6 short sentences) as part of the input. We hypothesize that the topic of a dialogue can be kept consistent, similar to how Wolf et al. [75] keep the personality of their chatbot consistent. The BFD Generator also uses only top-p sampling for decoding, unlike the TransferTransfo, which uses beam search with sampling. However, they are similar in the sense that multiple utterances are explored and ranked according to some metrics.

AdapterCL [47] AdapterCL concerns task-oriented dialogue, while the BFD Generator concerns conversational dialogue. AdapterCL trains each adapter on each domain (e.g., taxi or restaurant), while the BFD Generator trains an adapter on forward utterance generation, and an adapter on backward utterance generation. The BFD Generator does not have a classifier for which adapter to use (which may need to be trained when newer domains are added), but rather a static utterance scoring method. We hypothesize that by taking a similar mixture model approach [34], the independent adapters can optimize their individual performance on their specific tasks. If we train a single adapter on both forward and backward utterance generation, it may stumble upon the continuous learning problem of forgetting previous tasks.

LaMDA [71] The BFD Generator does not use human evaluations on dialogues for training. While both of the models generate multiple utterances, rank them, and select the best ranking utterance, the BFD Generator does this in separate modules. The BFD Generator generates utterances in its language model and ranks them in its Selection Module. LaMDA generates and ranks the utterances using only one language model. LaMDA also cooperates with a knowledge-retrieval system, while the BFD Generator does not, which can cause it to be lacking on informativeness and groundedness (see Section 2.6.1) in its generated utterances. Additionally, due to not being trained on a human-labelled dataset with labels for sensibleness (Section 2.6.1), the BFD Generator may be lacking on sensibleness. However, we hypothesize that its specificity, interestingness, and usefulness (see Section 2.6.1) will be high as a trade-off, since nonsensible utterances can prove useful as inspiration and brainstorming material.

Chapter 5

Experiments and Results

This chapter presents various experiments conducted to answer the RQs of this master's thesis (described in Section 1.3). First, all parameters and steps needed to reproduce the experiments are presented in Section 5.1. Each of the experiments are described in the following sections, Section 5.2, Section 5.3, Section 5.4, and Section 5.5, including the goal of the experiments (i.e., to answer the related RQ), research methods, and results along with discussions. Finally, the limitations of the experiments are discussed in Section 5.6.

5.1 Experimental Setup

This section presents the experimental setup by going through each part of the BFD Generator separately. This includes the tuning of the causal language model, the pre-processing of the dataset, and the decoding method. For each of the experiments, the values of the hyperparameters are as given in this section, unless otherwise specified.

Causal Language Model A model for forward generation is developed by adapter-based tuning a GPT-2 model using the `adapter-transformers` library [57]. It is tuned on the preprocessed Wizard of Wikipedia dataset, using Residual Adapters [32] on every layer. A model for backward generation is developed in similar fashion. In addition, two models (for forward and backward generation, respectively) are developed by fine-tuning GPT-2 on the same dataset using the `transformers` library [76]. The hyperparameters for all of the models can be seen in Table 5.1.

Parameter	**Value**	**Description**
η	$6.25 * 10^{-5}$	Learning rate of the Adam optimizer.
Gradient accumulation steps	4	Number of steps to accumulate over.
Number of distractors	1	Number of distractors for the next-sentence prediction head.
Maximum history	4	Number of the closest utterances to keep in the history.
$\|B\|$	4	Batch size for training and validation.
Number of epochs	2	Number of training epochs.

Table 5.1: The values of the hyperparameters for training the causal language model for the experiments.

Dataset The WIZARD OF WIKIPEDIA [20] is retrieved from the `ParlAI` framework [51] with the training, validation and test split by Dinan et al. [20]. We use only the dialogues labelled with topics.

The pre-processing of the dataset consists of simple data cleaning and removal of outliers. Augmentation of the dialogues is kept to a minimum to keep the integrity of the dialogues. Redundant whitespaces are removed from the dialogues. Dialogues with non-conversational or unintelligible data are outliers and are removed from the dataset. A dialogue is deemed an outlier if any of the following points apply.

- The dialogue contains utterances with image file extensions (i.e. ".png", ".svg", ".jpg", or ".jpeg"). Utterances with image file extensions often contain unintelligible data due to the data formats of images and/or text copied from an external website.

- The dialogue contains utterances with square brackets. This is because the square brackets often indicate a hyperlink to a website (e.g. "[Read more]") followed by unintelligible data due to the data formats of text copied from an external website.

- The dialogue contains at least one utterance with more than 200 tokens. This is due to memory limitations in the training process.

For generating the distractors, random utterances are picked from the pool of all utterances in the training set (as long as it does not equal the correct utterance). The seed is set to 24.

For random picking of dialogue snippets from the test set in Experiment 2, the seed is set to 24.

Decoding Method For the decoding method, the random seed is set to 24. Top-p sampling is used with p set to 0.9 with temperature θ set to 0.7. The number of candidates generated k is set to 8.

Selection Module For encoding sentences, we use the Universal Sentence Encoder [11] provided as a pipeline in the `spaCy` library.[1] For counting proper nouns, we use the `en_core_web_sm` language model with `spaCy` for encoding words, recognizing and removing stopwords, and recognizing proper nouns. The parameters of the Selection Module are given in Table 5.2.

Parameter	Value
g_{self}	0.42
g_{response}	0.43
k_1	0.1
k_2	5
k_3	4
k_4	0.01
k_5	0.1

Table 5.2: The values of the hyperparameters of the Selection Module for the experiments.

Content Creators We are provided two content creators from the communication exercise company. They have competence in education and content creation for communication exercises. Additionally, the content creators were chosen since they were interested in this thesis and wanted to volunteer. The content creators were not paid for their contributions to this thesis.

5.2 Experiment 1: Adapter-based Tuning

The goal of this experiment is to evaluate the benefits of adapter-based tuning versus fine-tuning in the context of:

RQ1 How can large-scale, general, deep learning language models be adapted to topic-specific surrounding dialogue generation in a scalable way using adapters?

[1] https://spacy.io/

Method We train adapters using the BFD Generator approach (see Section 4.3) on the Wizard of Wikipedia dataset with the experimental setup described in Section 5.1, resulting in an adapter for forward utterance generation and another adapter for backward utterance generation. Additionally, we fine-tune GPT-2 using the same dataset, experimental setup, and approach. This results in a fine-tuned language model for forward utterance generation and another for backward utterance generation. The main difference between the adapter-based tuning approach and the fine-tuning approach is which layers are trained. In the adapter-based tuning approach, only the adapters between all of the decoder-layers of GPT-2 are trained. In the fine-tuning approach, all of the decoder-layers of GPT-2 are trained. However, fine-tuning GPT-2 required more VRAM than available, thus, the batch size $|B|$ for fine-tuning is reduced to 1 (instead of 2). All of the models are trained on a single GeForce 980 Ti GPU with 6 VRAM. The training set for forward utterance generation is 80 MB and for backward utterance generation, 84 MB. Each of the datasets spans 1247 different topics, and contains 129.6k dialogue snippets, a total of 669.4k turns, and 259.2k distractors.

Results and Discussion The training time and required space for adapter-based tuning and fine-tuning are seen in Table 5.3. On a single GeForce 980 Ti GPU with 6 VRAM, it took about 4.9 hours to train the adapter for backward utterance generation and 4.8 hours to train the adapter for forward utterance generation. Fine-tuning GPT-2 for forward utterance generation took 15 hours, an 306% increase in training time compared to the forward utterance generation adapter. A similar increase is seen for backward utterance generation. By running some qualitative studies on input-output pairs, we observe that both of the models are able to generate utterances that are on-topic and possible to be interpreted by humans. Additionally, the storage space required for backward and forward utterance generation using fine-tuned language models (1000MB in total) is bigger than when using adapters (802MB in total, where each adapter takes 151MB and the small base GPT-2 model takes 500MB).

Thus, it is less expensive in both training time, VRAM requirements during training, and storage space to use adapter-based tuning instead of fine-tuning, while still gaining the ability to generate on-topic surrounding dialogue. This also applies to the addition of new adapters or swapping of existing experts of the BFD Generator. Thus, this experiment suggests that large-scale general language models (like GPT-2) can be adapted for topic-specific surrounding dialogue generation using the BFD Generator approach with adapter-based tuning, with scalability benefits in terms of training time, and storage space requirements.

Transfer learning technique	Expert	Training time	Space required
Adapter-based tuning	Backward	4.9 hours	151 MB
Adapter-based tuning	Forward	4.8 hours	151 MB
Fine-tuning	Backward	15 hours	500 MB
Fine-tuning	Forward	15 hours	500 MB
-	GPT-2	-	500 MB

Table 5.3: The training time and space required for the models using adapter-based tuning and fine-tuning. The total size of the language model for backward and forward generation is 802 MB (including the base GPT-2 model). The total size of the fine-tuned equivalent is 1000 MB.

5.3 Experiment 2: Scoring and Selecting Utterances

The goal of this experiment is to evaluate the ability of our utterance scoring method (i.e., the Selection Module) to select among a pool of BFD-generated utterances a suitable utterance to the dialogue snippet. A suitable utterance is an utterance a content creator would choose given the same options. For a content creator, many factors may influence their choice. For instance, a suitable utterance may be the one that keeps the flow of a conversation, however, it is up to them to decide. Thus, we research:

RQ2 How do data-driven approaches, for selecting generated preceding and following utterances to add to a dialogue snippet, align with human content creators' judgements?

Method We send a questionnaire to two content creators containing forty multiple-choice tasks. The exact instructions and tasks sent to the content creators are in Appendix A.1. In each task (see Figure 5.1), the content creators are presented a dialogue snippet, its topic and CEFR level, and are asked to select the best utterance(s) out of eight options. The dialogue snippets (and their topics) are randomly picked from the test set in the WIZARD OF WIKIPEDIA dataset [20], each cut to a random length of at least one turn. A random CEFR level is attached to each task. The options are generated by the adapter-based tuned language model of the BFD Generator. In half of the tasks, the content creator is asked to pick the best forward candidate(s). In the rest, they are asked to pick the best backward candidate(s). The reason for separating these tasks is to make the form easier for the content creators to understand and finish.

From the data, we compute the accuracy of variants of Selection Module's scoring method and compare them to a baseline. Let the best candidates for

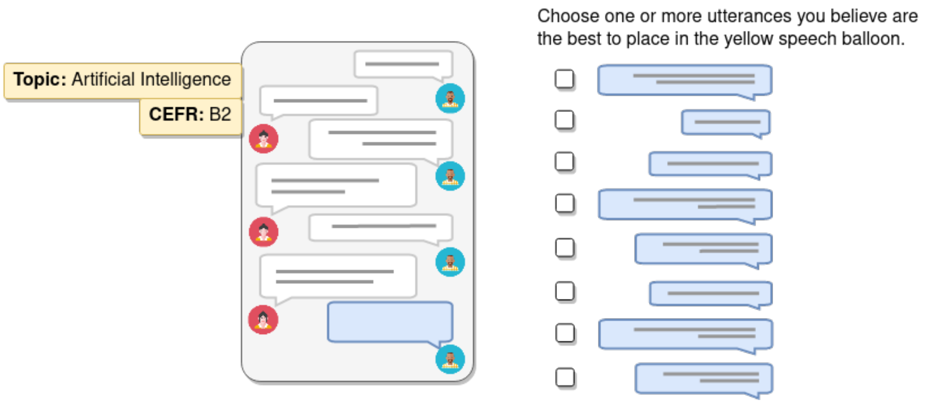


Figure 5.1: In Experiment 2, a questionnaire consisting of forty multiple-choice tasks (like illustrated above) is given to the content creators. In each task, a prompt (to the left) consisting of a dialogue, its topic and a random CEFR level is shown. In addition, the content creator is asked to choose one or more utterances (to the right) they believe the best to add to the dialogue in the blue speech balloon (at the bottom to the left).

each dialogue snippet in the questionnaire be the union of the content creators' picks. Let T be the number of tasks where the highest-scoring utterance (by the module) is one of the best candidates. Let N be the number of tasks. Thus, the accuracy is:

$$\alpha = T/N, \tag{5.1}$$

which represents the module's performance to select a candidate like a human expert would. If the agent's accuracy α is 100%, then it always picks an utterance accepted by a human. However, this does not tell anything about how well the rest of the candidates are ranked.

Thus, we also study the top-b accuracy β, which represents the module's performance to have an overlapping understanding of acceptable candidates with human experts. Let c_i be the number of candidates that are chosen and are also the best candidates for the task $i \in [0, T]$. Let b_i be the number of best candidates for task i. The top-b accuracy is:

$$\beta = \frac{\Sigma_{i=0}^{N} \frac{c_i}{b_i}}{N}. \tag{5.2}$$

If an agent's top-b accuracy is 100%, then it always picks the same utterances as the humans, given it was allowed to pick as many as them. This indicates that the agent is able to rank the utterances such that the most suitable ones are amongst the top, though it may not necessarily be good at picking out the best among the best.

Results and Discussion The responses to the questionnaire are shown in Appendix A.2. Table 5.4 shows the result of this study of the Selection Module's scoring function and its performance compared to a baseline. The baseline is computed as the average accuracies of random scoring modules. A random scoring module gives each candidate a score by drawing a variable X from the standard normal distribution $\mathcal{N}(0,1)$, i.e., $X \sim \mathcal{N}(0,1)$. The baseline is then computed as the averaged accuracies over the seeds $i = 0, ..., S$:

$$\begin{aligned} \alpha_{\text{baseline}} &= \sum_{i=0}^{S} \alpha_i / S, \\ \beta_{\text{baseline}} &= \sum_{i=0}^{S} \beta_i / S, \end{aligned} \tag{5.3}$$

where S is 999 and α_i and β_i are the accuracies of the random scoring module on the seed i.

In Table 5.4, we observe that all of the individual components of the candidate scoring function are improvements from the baseline, with peaks in the individual dissimilarity scores d_{self} and d_{response}. Self dissimilarity acquires the highest β accuracy, suggesting that it may be important to keep the scores of the best b candidates (chosen by human content creators) the highest among the candidates. Response dissimilarity acquires the highest α accuracy, suggesting it may be important to select a candidate accepted by a human content creator.

While the scoring method we use in the BFD Generator does not acquire the highest accuracies (α_{BFD} and β_{BFD}), its accuracies are among the highest in total. Thus, preceding and following utterances can be added to a dialogue snippet by scoring the utterances using proper noun count and dissimilarity scores as an improvement to assigning them random scores (simulating picking a random utterance, or just picking the first utterance generated). However, all of the scoring methods disagree with the human content creators' judgements more than 50% of the time. Also, we only achieve less than 7% higher accuracies using our scoring method compared to the baseline. This suggests that utterance scoring and selection are still challenging research areas.

$s(u_c)$	**Accuracy** α	**Top-b accuracy** β
n_c	0.3500	0.3516
d_{self}	0.4000	**0.4207**
d_{response}	**0.4500**	0.3680
$k_2 * d_{\text{self}} + k_3 * d_{\text{response}}$	0.3750	0.3807
$k_1 * n_c + k_2 * d_{\text{self}} + k_3 * d_{\text{response}}$	$\alpha_{\text{BFD}} = 0.4000$	$\beta_{\text{BFD}} = 0.3974$
Baseline	$\alpha_{\text{baseline}} = 0.3312$	$\beta_{\text{baseline}} = 0.3316$

Table 5.4: The accuracy and top-b accuracy of various scoring functions $s(u_c)$ for the Selection Module and a baseline computed from the average of random scoring modules. The BFD Generator uses the scoring function in the penultimate row, with the accuracies α_{BFD} and β_{BFD}.

5.4 Experiment 3: Content Creator User Study

The goal of this experiment is to evaluate the effectiveness of the BFD Generator as a creative tool for content creators in the education domain through a quantitative and qualitative study of a user study. Specifically, we seek to answer the following RQ:

RQ3 Do extended dialogues through data-driven approaches, after adjustments made by experts, result in high-quality dialogues appropriate as specialized content?

Method The user study is divided into four parts (similar to Figure 1.2):

1. Firstly, content creators are asked to create dialogue snippets that we generate dialogues from using variants of the BFD Generator (see Figure 5.2). A surrounding dialogue for each CEFR level is generated, except for the B2 level where we generate two dialogues instead. This is because the BFD Generator is trained on mostly dialogues on a B2 level. Thus, better-extended dialogues may be generated on this level. *BFD-*, *BD-*, and *FD-extended dialogues* refer to the dialogues extended using the BFD Generator, the BD Generator, and the FD Generator, respectively (see Figure 5.2).

2. Secondly, the content creators evaluate the *extended dialogues* (without any adjustments) using the gold-standard human evaluation metrics (see Section 2.6.1). In addition, they are asked how well the generated dialogue works as a dialogue exercise for learning to speak about the topic on the specific CEFR level. The evaluations are given on a 7-point Likert scale.

3. Thirdly, the content creators are given five minutes to make adjustments to the generated draft extended dialogues to make them suitable as an exercise for teaching to speak about the topic on the specific CEFR level. We refer to these dialogues as *adjusted dialogues*. Adjusted BFD-, BD-, and FD-extended dialogues refer to the BFD-, BD-, and FD-extended dialogues *adjusted* by content creators.

4. Lastly, the content creators evaluate the edited dialogues on the same metrics as in the second step. They are in addition asked how useful the generated dialogue was for improving the dialogue exercise creation experience (also given on a 7-point Likert scale).

The exact instructions to the content creators are in Appendix A.3. We conduct the user study with two content creators (separately). The reason for getting the opinions from multiple content creators is due to the creative nature of the content creation task, adding subjectivity to the evaluations.

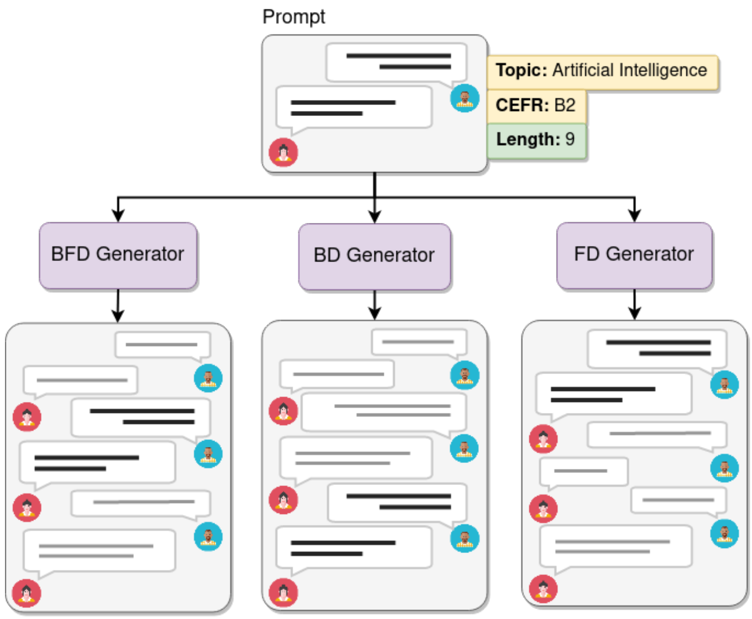
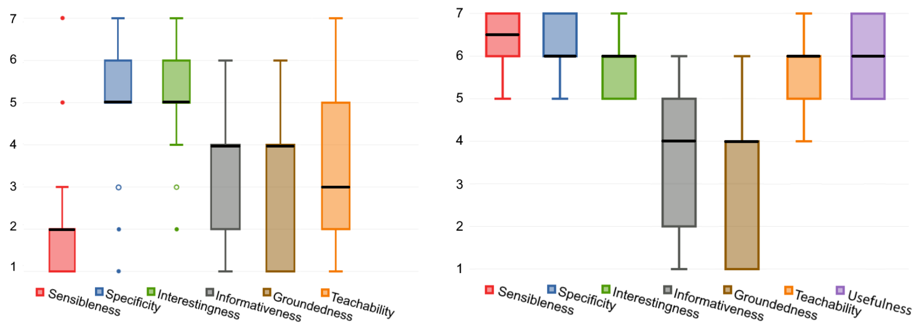


Figure 5.2: A prompt (on the top) consisting of a topic (e.g., Artificial Intelligence), a dialogue snippet on that topic (indicated by black lines in text bubbles), and its CEFR level (e.g., B2) is extended with forward and backward utterances (grey lines) using variants of the BFD Generator. The BFD Generator (to the left) generates both preceding and following utterances. The BD Generator (in the middle) only generates preceding utterances. The FD Generator (to the right) only generates following utterances. All of the generated extended dialogues consist of nine turns.

5.4.1 Results and Discussion

We only discuss some of the most important results for **RQ3** and **RQ4** in this section. However, the rest of the results are available in Appendix A.4.

Quantitative Study Boxplots for results of the user study are shown in Figure 5.3. From the results, we see that the content creators had no difficulties increasing the sensibleness, interestingness and teachability of the dialogues consistently. Notably, the content creators consistently add significant value to sensibleness and teachability. However, we observe a big interquartile range in informativeness and groundedness, both in the generated and human adjusted dialogues. We hypothesize that these two metrics are deemed not important for our purposes or the time limit was too tight. Given limited time, less attention was given to improving informativeness and groundedness. Throughout the entire study, the generated dialogues (with only forward, only backward and both utterances) were useful to the content creators to make language learning content as shown by the boxplot of usefulness (in purple to the right in Figure 5.3b).

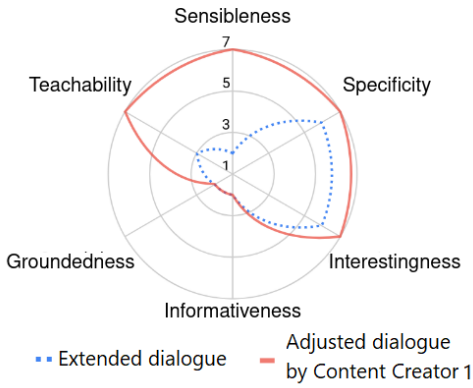


(a) Extended dialogues (b) Adjusted dialogues

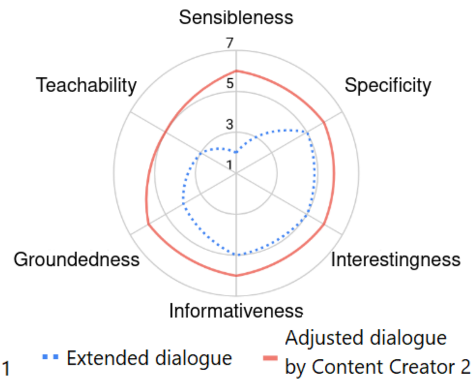
Figure 5.3: Boxplot of human evaluations of extended dialogues before (5.3a) and after (5.3b) the content creators make dialogue adjustments. The content creators score the extended dialogues and adjusted dialogues on the six metrics (on the horizontal axis) on a 7-point Likert scale, where 1 is the worst, 4 is neutral and 7 is the best score. Feedback on the usefulness (in purple to the right in 5.3b) of the extended dialogues for creating content on the specific topic and CEFR level is given on the same scale.

When comparing the evaluations of the BFD-extended dialogues (see Figure 5.7) and self-evaluations of the adjusted dialogues (see Figure 5.8), both of the content creators experienced an increase across different metrics with only five minutes to make adjustments. While both content creators managed to improve the teachability, sensibleness, specificity, and interestingness of the dialogues, there was little increase in groundedness and informativeness. This suggests that the task of improving groundedness and informativeness may be more time-consuming or considered less of a priority for content creators than improving the dialogues in the other metrics.

Qualitative Study We study how the content creators made use of the BFD Generator for an example from the user study. Table 5.5 shows a BFD-extended dialogue and the content creators' versions of it after being given five minutes



(a) Evaluations by Content Creator 1



(b) Evaluations by Content Creator 2

Figure 5.4: The content creators' individual evaluations of the dialogues in Table 5.5 differ most notably on groundedness and informativeness, but are similar in teachability, sensibleness, specificity, and interestingness. Both content creators increased the teachability, sensibleness, and specificity of the dialogues in less than five minutes.

to adjust it. Their evaluations and the improvements of the dialogues across the metrics are seen in Figure 5.4.

We observe that both of the adjusted dialogues use parts of the generated utterances, notably more in the preceding utterances to the dialogue snippet. They found a phrase that suited the CEFR level (i.e., "I love traveling" on line 2 in Table 5.5) and used repetition to make it more suitable as part of a language exercise. In the forward utterances of the dialogue snippet, there is less resemblance between the generated dialogue and the content creators' versions. It is clear that the generated forward utterances do not fit as well with the dialogue snippet, thus, the content creators themselves increase the quality of the dialogue.

An interesting find is how Content Creator 1 uses the uncertainty in the generated utterance "I don't really know much about travel, but I know that I love to travel" (see line 8 in Table 5.5) by changing the reasoning behind the uncertainty to make it fit the context better (see line 17 in Table 5.5). Another interesting point is how the adjusted dialogues are similar in nature, starting with expressing interest for travelling (see line 10-11, and 20-21 in Table 5.5) and ending with recommendations for attractions (see line 18 and 26 in Table 5.5).

The evaluations of the groundedness and informativeness of the dialogues are significantly different between the content creators. However, the content creators' scores for the sensibleness, specificity, interestingness, and teachability of the generated dialogue align well with each other. This shows the subjectivity of human judgements. This example suggests that the content creators are able to productively improve the BFD-generated output, even though it is of lower quality along with some metrics.

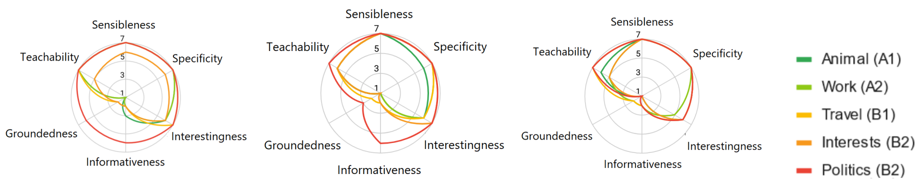
Line	BFD-extended Dialogue on the topic "Travel" and CEFR level B1
1	A: I have! What is your favorite travel destination?
2	B: I love traveling. I love to travel, do you?
3	A: Yes, I've been to Greece. Have you been to Greece?
4	B: **What are your plans for the summer holiday?**
5	A: **We are going to Greece. I can't wait!**
6	B: **That's great, I love Greece! Have you been there before?**
7	A: **No, it's my first time. Do you have any recommendations?**
8	B: I don't really know much about travel, but I know that I love to travel.
9	A: I have heard that travel is one of the most important activities for the human race. Do you know if that is true?
	Content Creator 1's Adjusted Dialogue
10	A: I love to travel, do you?
11	B: Yes, I love traveling. I'm going to Italy this summer.
12	A: That sounds nice!
13	B: **What are your plans for the summer holiday?**
14	A: **We are going to Greece. I can't wait!**
15	B: **That's great, I love Greece! Have you been there before?**
16	A: **No, it's my first time. Do you have any recommendations?**
17	B: I haven't been in Greece, so I don't know. Have you been in Italy and can give me some recommendations?
18	A: Yes, I have been in Italy. You should visit the Colosseum in Rome!
	Content Creator 2's Adjusted Dialogue
19	A: Do you like to travel?
20	B: I love traveling. I love to travel, do you?
21	A: Yes, I love to travel as well.
22	B: **What are your plans for the summer holiday?**
23	A: **We are going to Greece. I can't wait!**
24	B: **That's great, I love Greece! Have you been there before?**
25	A: **No, it's my first time. Do you have any recommendations?**
26	B: I think the Parthenon in Athens is amazing.
27	A: Then we will definitely go there. Do you have any other recommendations?

Table 5.5: A dialogue is BFD-extended (on the top) by using the input dialogue snippet (in bold) and the input topic "Travel". Each utterance is labelled with a line number (left column). The content creators are given five minutes to make adjustments to the generated dialogue to make it a suitable communication string exercise for the CEFR level B1 given the topic. The content creators were informed that their final dialogues (at the bottom) must contain the dialogue snippet.

Surrounding Dialogue vs One-Directional Generation To further discuss the value of surrounding dialogue generation in the context of **RQ3**, we compare the self-evaluations of the adjusted BFD-extended dialogues with adjusted BD- and FD-extended dialogues. An important note to take into consideration is that in the user study, the content creators adjusted BFD-extended dialogues first, then BD-generated, and finally FD-generated. Thus, there is a possibility that the results may be affected by how familiar the tasks in the user study become to the content creators.

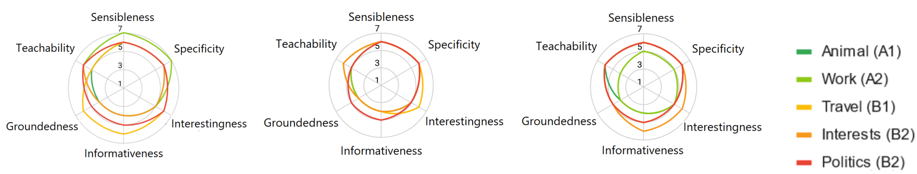
In Figure 5.5 and Figure 5.6, self-evaluations of all adjusted BFD-, BD-, and FD-extended dialogues by the content creators are shown. The first observation we make is that the overall shape is similar between adjusted BFD-, BD-, and FD-extended dialogues when looking at the evaluations from the content creators separately. In general, we observe that adjusted BFD-extended dialogues for both of the content creators score the highest across all metrics. This suggests that while both preceding utterances (represented through adjusted BD-extended dialogues) and following utterances (represented through adjusted FD-extended dialogues) are useful separately, it is together that the best resulting extended dialogues may be achieved the fastest.

Additionally, the content creators told us that all of the extended dialogues were on-topic. This supports our hypothesis about being able to keep the topic of the dialogue consistent by adapting a transfer learning technique for keeping the personality of the speakers consistent. The content creators also found the extended dialogues to be on a too-high CEFR level in most cases. This may be due to the CEFR imbalance in the dataset.



(a) BFD-extended (b) BD-extended (c) FD-extended

Figure 5.5: Evaluations of all adjusted BFD-, BD-, and FD-extended dialogues by Content Creator 1.



(a) BFD-extended (b) BD-extended (c) FD-extended

Figure 5.6: Evaluations of all adjusted BFD-, BD-, and FD-extended dialogues by Content Creator 2.

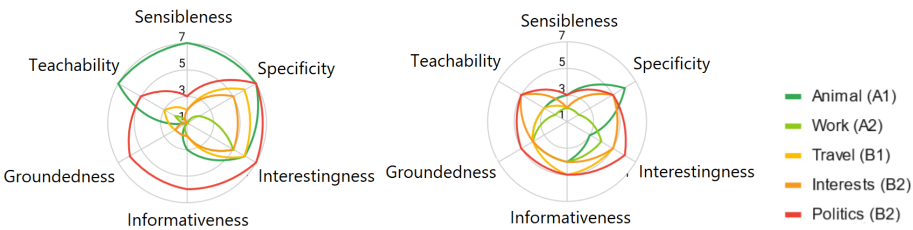
5.5 Experiment 4: Comparison of Content Creators

The goal of this experiment is to compare the evaluations of the content creators, in the context of the following RQ:

RQ4 How does the subjective and open-ended nature of content creation impact evaluations of extended and adjusted dialogues?

Method We pick relevant results from the user study in Experiment 3 (Section 5.4).

Results and Discussion We compare the two content creators' evaluations of BFD-extended dialogues (see Figure 5.7) and self-evaluations of the adjusted BFD-extended dialogues (see Figure 5.8).

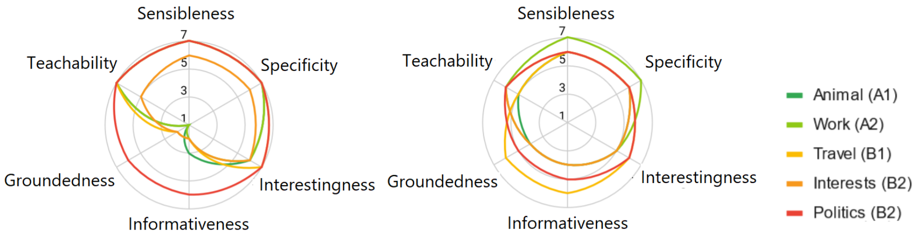


(a) Evaluations by Content Creator 1

(b) Evaluations by Content Creator 2

Figure 5.7: The content creators' individual evaluations of all BFD-extended dialogues. In the radar charts, each color is the evaluation of a single dialogue on its respective topic and CEFR level (to the right). Their evaluations differ mostly in groundedness and informativeness and in the dialogue about animals on A1 level.

First, we observe how the content creators' evaluations vary when given the same dialogues in Figure 5.7. Generally, the content creators score the BFD-extended dialogues low on sensibleness and high on specificity. However, Content Creator 1 scores the dialogues significantly lower on groundedness and informativeness than Content Creator 2 in general (with a single dialogue as an exception). This suggests that the content creators have different requirements for groundedness and informativeness. Content Creator 1 also has more spikes in their evaluations (see the two outliers in dark-green and red in Figure 5.7a). This may not only be a result of their subjective interpretations of the dialogues, but



(a) Self-evaluations by Content Creator 1

(b) Self-evaluations by Content Creator 2

Figure 5.8: The content creators' self-evaluations of all adjusted BFD-extended dialogues. In the radar charts, each color is the evaluation of a single dialogue on its respective topic and CEFR level (to the right). Generally, the adjusted BFD-extended dialogues by Content Creator 1 score low on groundedness and informativeness and high on the rest. On the other hand, Content Creator 2 score medium to high across all metrics.

also the discovery of useful or inspirational fragments of the dialogue. In particular, Content Creator 1 commented on the generated extended dialogue (see the dark-green in Figure 5.7a) on the topic "Animal" and CEFR level A1: "It is good teaching material because it covers basics of A1 such as basic verbs, negation, question words and being able to have a conversation about personal interests". Since the content creator recognized and valued these elements, they scored the generated extended dialogue high on teachability. Thus, the evaluations of the teachability of dialogues can be strongly affected by the content creator's ability to define and recognize important elements.

Secondly, we look into how the resulting adjusted BFD-extended dialogues differ in Figure 5.8. Due to their limited time to make adjustments to the dialogues, the content creators must use their time wisely. This may reveal what aspects of the dialogues are the most important ones for the content creators and aspects that may be more time-consuming to improve if the evaluations of the adjusted dialogues (in Figure 5.8) differ between the content creators. The adjusted dialogues differ mostly in teachability, groundedness, and informativeness. In Figure 5.8a, Content Creator 1 perceives that their adjusted dialogues score high in teachability, sensibleness, specificity, and interestingness, but lack in groundedness and informativeness in general. In contrast, Content Creator 2 (in Figure 5.8b) perceives that their adjusted dialogues are balanced across all metrics.

The evaluations and self-evaluations by the content creators differ significantly on multiple metrics, and that is with a trend. Thus, in the context of

RQ4, subjectivity affects the evaluations so significantly the evaluations should be analyzed separately to understand the gains the content creators experience from the extended to adjusted dialogues.

5.6 Limitations

We have made several limitations to our experiments for feasibility, given our resources. In this section, we discuss the limitations of our experiments.

To make the user studies feasible, given our resources, we made several limitations to them. First, we use small sample sizes in our user studies, both in terms of dialogues and participants. In Experiment 2 (see Section 5.3), we collect evaluation data for forty dialogues, and in Experiment 3 (see Section 5.4), for thirty dialogues from two content creators. This makes the results of the user studies more vulnerable to inconsistencies caused by the subjective opinions of the participants and their individual understanding of each metric. On the other hand, it allows us to focus on how our architecture benefits specific content creators and how the benefits may vary. Through continuous communication with the content creators we collaborate with, we can gain qualitative insight into their experiences of our architecture.

Another limitation to the user study of Experiment 3 is that the content creators gain experience over time in solving the tasks we give them. The content creators might become more efficient at solving the tasks the more dialogues they see, evaluate and adjust. Combined with the time limit of the user study, the order we show the extended dialogues to the content creators may impact the resulting adjusted dialogues and cause inconsistencies in their judgements. While our user study demonstrates the benefits of the BFD Generator, a bigger user study with more content creators, where each of them solves a few tasks, is necessary to get a better understanding of the BFD Generator's capabilities.

When comparing fine-tuning to adapter-based tuning in Experiment 1 (see Section 5.2), we only conduct a simple evaluation of their text generation capabilities. To have an in-depth evaluation of the performance of the fine-tuning approach, a user study similar to the one conducted in Experiment 3 (see Section 5.4) should be done. However, user studies are expensive, and we did not have enough resources to conduct more user studies.

Chapter 6

Conclusion and Future Work

This chapter concludes this thesis in Section 6.1 by discussing each RQ proposed in Chapter 1 and hypotheses in Chapter 3. We propose future work in Section 6.2.

6.1 Conclusion

In this thesis, we present the BFD Generator, our solution to the novel SDGP. We take inspiration from both SOTA conversational and task-oriented dialogue systems, as well as efficient transfer learning techniques. Through various experiments, we answer all of the RQs (in Chapter 1) and hypotheses (in Chapter 3). In this section, we give a summary of our findings.

Hypothesis 1 Preceding utterances to a dialogue snippet can be generated using transfer learning techniques similar to those used in response generation systems for following utterances.

We adapt GPT-2 using an adjusted transfer learning technique inspired by a SOTA conversational dialogue system training approach, TransferTransfo [75], and a modular task-oriented dialogue system using adapter-based tuning [32, 6], AdapterCL [47]. We use adapter-based tuning to train two independent adapters for forward utterance generation and backward utterance generation. Our architecture is capable of generating surrounding dialogue, including preceding utterances to a dialogue snippet, supporting **Hypothesis 1**.

Hypothesis 3 An adapter can be assigned to each subtask of the SDGP to develop an efficient language model for surrounding dialogue generation.

RQ1 How can large-scale, general, deep learning language models be adapted to topic-specific surrounding dialogue generation in a scalable way using adapters?

In our comparison between our adapter-based tuning approach and an equivalent fine-tuning approach, we find that our approach is more efficient in terms of training time, storage space, and memory for training. However, all of the models are capable of generating on-topic preceding and following utterances to a dialogue snippet. Thus, our approach is more scalable, as new and improved adapters (e.g., with better performance, or concerning a new domain) can be trained using fewer resources and added to the BFD Generator, compared to using fine-tuning.

RQ2 How do data-driven approaches, for selecting generated preceding and following utterances to add to a dialogue snippet, align with human content creators' judgements?

The BFD Generator adds a preceding or following utterance to a dialogue snippet by generating multiple candidate utterances, similar to LaMDA [71], scoring them, and adding the best scoring candidate to the snippet. We present two metrics for the accuracy of our utterance scoring and selection method, accuracy α and top-b accuracy β. Our scoring method uses a weighted sum of noun phrase count, self dissimilarity score, and response dissimilarity score. Compared to scoring the utterances randomly, we observe a slight increase in performance using only proper noun count, and a higher increase using the dissimilarity scores individually. Self dissimilarity score achieves the highest top-b accuracy (β), suggesting that it may be important to keep the scores of the best b candidates (chosen by human content creators) the highest among the candidates. Response dissimilarity score achieves the highest accuracy (α), suggesting it may be important to select a candidate accepted by a human content creator. While the BFD Generator's scoring method does not acquire the highest accuracies, its accuracies are among the highest in total.

However, all of the scoring methods we tested did not exceed an improvement of 7% to the baseline. The methods also disagree with the human content creators' judgements in more than 50% of the tasks. This suggests that utterance scoring and selection using data-driven methods are still challenging tasks.

RQ3 Do extended dialogues through data-driven approaches, after adjustments made by experts, result in high-quality dialogues appropriate as specialized content?

Through user studies with content creators, we find that the content creators can increase the sensibleness, specificity, interestingness, and teachability of BFD-extended dialogues within only five minutes. The most significant improvements are seen in sensibleness and teachability. However, there is no noticeable consistent increase in informativeness and groundedness. In general, the content creators found the extended dialogues useful to create content to teach a person how to speak about a topic on a specific CEFR level.

In a comparison of surrounding dialogue generation to one-directional dialogue generation, we observe that dialogues extended with both preceding and following utterances score the highest across all metrics in general. This is despite having the content creators adjust the BFD extended dialogues before dialogues extended in only one direction.

RQ4 How does the subjective and open-ended nature of content creation impact evaluations of extended and adjusted dialogues?

How the content creators use extended dialogues to create content varies between them. Sometimes they keep the same utterances. Other times, they can be inspired by specific parts of the extended dialogues. This highly impacted their evaluations of the extended dialogues, as their perception of the value of the extended dialogues differed depending on what they could observe. In particular, this impacted their evaluation of the teachability of extended dialogues significantly. We also observe a trend in their evaluations of informativeness and groundedness. Specifically, one of the content creators scored the dialogues, both the extended and adjusted, lower in general than the other content creator. This suggests that the evaluation of informativeness and groundedness may be impacted by the content creators' subjective opinions. The subjective nature of content creation is so significant that their evaluations must be analyzed separately to understand the gains the content creators experience from the extended to adjusted dialogues.

Hypothesis 2 The evaluation metrics used in LaMDA can be used to evaluate dialogues by transforming them from utterance-level to dialogue-level and result in high-quality evaluations for sensibleness, specificity, and interestingness, but may be lacking in groundedness and informativeness.

The metrics we use (described in Section 2.6.1) in our user studies are based on the metrics used to evaluate LaMDA [71]. From the previous discussion, we find that these metrics are effective to evaluate different aspects of extended dialogues and adjusted dialogues. The metrics have allowed us to find the benefits and weaknesses of our SDGP architecture, and answer multiple RQs and hypotheses.

Goal Develop an efficient data-driven architecture which generates, from a dialogue snippet including topic-specific terminology, surrounding topic-specific dialogue suitable for communication exercises for humans after adjustments made by an expert.

In this thesis, we propose an efficient SDGP architecture capable of extending dialogue snippets with topic-specific preceding and following utterances. Through user studies, we observe that our architecture is useful for content creation when used as a creative tool by content creators. The content creators we worked with managed to create communication exercises of satisfactory quality, only spending five minutes to make adjustments to the machine-extended dialogues. Additionally, we find that our architecture using adapter-based tuning is efficient in terms of training time, storage space, and memory used during training, compared to fine-tuning. Thus, we have met the goal of this thesis.

6.2 Future Work

While we have covered several interesting research questions and hypotheses, there are still many interesting related areas to research. In this section, we propose some future work we find interesting, to either improve or give more insight into our architecture.

We suggest developing a dialogue system that can either generate or rephrase utterances to be on a specified CEFR level. The content creators that participated in our experiments informed us that most of the utterances generated by the BFD Generator are on the B2 level or higher. This required the content creators to rephrase the utterances to the appropriate CEFR level themselves. We suggest taking a data-driven approach, which needs a CEFR labelled dialogue dataset with enough samples for each CEFR level.

It is of interest to collect more evaluation data of the BFD Generator as a creative tool for content creation. Due to the complexity of evaluating the BFD Generator and its outputs, the lack of well-explaining automatic metrics for NLP tasks, and the resources available, limitations had to be imposed in experiments. While human judgements, especially experts', provide the most reliable feedback on the quality of the dialogues, collecting more data points could give us more insight. It would be interesting to cover more dialogues over various topics, of different lengths and on different CEFR levels. However, this is expensive, so this is an inherent trade-off.

We propose an experiment where surrounding dialogues written without any generative tools are evaluated. This is to further show the benefits of the BFD Generator. We suggest experimenting using the method of Experiment 3 (see Section 5.4), but with only the dialogue snippets (and none of the generated utterances). This allows for a direct comparison of dialogues written with (using our results) and without the BFD Generator.

The development of larger language models has brought interest in few-shot learning as a possible transfer learning technique for adapting models [9]. A comparison of few-shot learning with our adapter-based tuning approach can give us insight into the performance, scalability, and sustainability of the approaches. Depending on the results, one can adapt the BFD Generator to use the most efficient transfer learning approach.

We are also interested in improving the performance of our utterance scoring technique. We suggest adapting a language model to score utterances. The training set can consist of dialogue snippets, each with multiple possible preceding and following utterances labelled with scores. The language model's task is to predict the score of each utterance, given the dialogue snippet. The utterances can be scored by crowd workers, much like how data was collected for LaMDA [71].

Finally, we want to mention that our research has not focused on the safety of the generated surrounding dialogue, even though this is an important aspect to consider. Thus, we suggest researching the prevention of inappropriate utterance generation. This is because the BFD Generator uses GPT-2, which is trained on human-derived data. Even with a content creator evaluating the BFD-extended dialogues, a mistake may result in the distribution of unethical dialogues. Additionally, the content creators should preferably not have to be subjected to inappropriate utterances.

Bibliography

[1] Ahmed R. Abas, Ibrahim El-Henawy, Hossam Mohamed, and Amr Abdellatif. Deep Learning Model for Fine-Grained Aspect-Based Opinion Mining. *IEEE Access*, 8:128845–128855, 2020. doi: 10.1109/ACCESS.2020.3008824.

[2] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for Document Classification. *CoRR*, abs/1904.08398, 2019. URL `http://arxiv.org/abs/1904.08398`.

[3] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a Human-like Open-Domain Chatbot. *CoRR*, abs/2001.09977, 2020. URL `https://arxiv.org/abs/2001.09977`.

[4] Obinna Agbodike, Chiao-Hua Huang, and Jenhui Chen. Cognitive Attention Network (CAN) for Text and Image Multimodal Visual Dialog Systems. In *International Conference on Applied System Innovation (ICASI)*, pages 37–41, 2020. doi: 10.1109/ICASI49664.2020.9426334.

[5] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint Language and Translation Modeling with Recurrent Neural Networks. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1044–1054, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1106`.

[6] Ankur Bapna and Orhan Firat. Simple, Scalable Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1165. URL `https://aclanthology.org/D19-1165`.

[7] Hannah Bast, Matthias Hertel, and Mostafa M. Mohamed. Tokenization Repair in the Presence of Spelling Errors. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 279–289, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.conll-1.22.

[8] Will Bridewell and Paul Bello. A Theory of Attention for Cognitive Systems. *Advances in Cognitive Systems*, 4(1):1–16, 2016.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[10] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547.

[11] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029. URL https://aclanthology.org/D18-2029.

[12] Ellen Zhang Chang. Specialization Project Report. In *TDT4501*. Norwegian University of Science and Technology, 2021.

[13] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George F. Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. The Best of Both Worlds:

Combining Recent Advances in Neural Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*, 2018.

[14] Zhiyu Chen, Honglei Liu, Hu Xu, Seungwhan Moon, Hao Zhou, and Bing Liu. NUANCED: Natural Utterance Annotation for Nuanced Conversation with Estimated Distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4016–4024, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.337. URL `https://aclanthology.org/2021.findings-emnlp.337`.

[15] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1053. URL `https://aclanthology.org/D16-1053`.

[16] Hamid R. Chinaei and Brahim Chaib-draa. *Building Dialogue POMDPs from Expert Dialogues - An end-to-end Approach.* Springer International Publishing, 2016. ISBN 978-3-319-26200-0. doi: 10.1007/978-3-319-26200-0.

[17] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, 54:755 – 810, 2020.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[19] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. The Second Conversational Intelligence Challenge (ConvAI2). *CoRR*, abs/1902.00098, 2019. URL `http://arxiv.org/abs/1902.00098`.

[20] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[21] Fatima Zahra Fagroud, Mohamed Rachdi, and El Habib Ben Lahmar. Automatic Story Generation: Case Study of English Children's Story Generation

Using GPT-2. In *Digital Technologies and Applications*, pages 54–62, Cham, 2022. Springer International Publishing. ISBN 978-3-031-02447-4.

[22] Yao Fu, Yansong Feng, and John P. Cunningham. Paraphrase Generation with Latent Bag of Words. In *Advances in Neural Information Processing Systems*, 2019. URL `https://proceedings.neurips.cc/paper/2019/file/5e2b66750529d8ae895ad2591118466f-Paper.pdf`.

[23] Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. Generating Distractors for Reading Comprehension Questions from Real Examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[24] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1443. URL `https://aclanthology.org/D18-1443`.

[25] Venkat N. Gudivada, Dhana Rao, and Vijay V. Raghavan. Chapter 9 - Big Data Driven Natural Language Processing Research and Applications. In *Big Data Analytics*, volume 33 of *Handbook of Statistics*, pages 203–238. Elsevier, 2015. doi: https://doi.org/10.1016/B978-0-444-63492-4.00009-5. URL `https://www.sciencedirect.com/science/article/pii/B9780444634924000095`.

[26] Hani Hagras. Toward Human-Understandable, Explainable AI. *Computer*, 51:28–36, 2018.

[27] J.B. Hampshire and A. Waibel. The Meta-Pi Network: Building Distributed Knowledge Representations for Robust Multisource Pattern Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7): 751–769, 1992. doi: 10.1109/34.142911.

[28] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2208–2222, 2021. URL `https://doi.org/10.18653/v1/2021.acl-long.172`.

[29] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-Based Sentiment Analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland, 2019. Linköping University Electronic Press. URL `https://aclanthology.org/W19-6120`.

[30] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

[31] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/e946209592563be0f01c844ab2170f0c-Paper.pdf`.

[32] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019. URL `https://proceedings.mlr.press/v97/houlsby19a.html`.

[33] Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1252`.

[34] Robert A. Jacobs and Michael I. Jordan. Learning Piecewise Control Strategies in a Modular Neural Network Architecture. *IEEE Trans. Syst. Man Cybern.*, 23:337–345, 1993.

[35] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1): 79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

[36] Rosalie Kern, Terry Libkuman, Hajime Otani, and Katherine Holmes. Emotional Stimuli, Divided Attention, and Memory. *Emotion (Washington, D.C.)*, 5:408–17, 2006. doi: 10.1037/1528-3542.5.4.408.

[37] Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Empirical Methods in Natural Language Processing*, 2020.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in*

Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[39] Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. Large-Scale Contextualised Language Modelling for Norwegian. *CoRR*, abs/2104.06546, 2021. URL https://arxiv.org/abs/2104.06546.

[40] Jiwei Li and Dan Jurafsky. Mutual Information and Diverse Decoding Improve Neural Machine Translation. *CoRR*, abs/1601.00372, 2016. URL http://arxiv.org/abs/1601.00372.

[41] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1099.

[42] Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. Corpora Generation for Grammatical Error Correction. 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 3291–3301, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1333. URL https://aclanthology.org/N19-1333.

[43] Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1187. URL https://aclanthology.org/N18-1187.

[44] Bing Liu, Tong Yu, Ian R. Lane, and Ole Jakob Mengshoel. Customized Nonlinear Bandits for Online Response Selection in Neural Conversation Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[45] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1116–1126, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1103. URL `https://aclanthology.org/P17-1103`.

[46] Xiren Ma and Azzedine Boukerche. An AI-based Visual Attention Model for Vehicle Make and Model Recognition. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6, 2020. doi: 10.1109/ISCC50000.2020.9219660.

[47] Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. Continual Learning in Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.590. URL `https://aclanthology.org/2021.emnlp-main.590`.

[48] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL `https://www.sciencedirect.com/science/article/pii/S0079742108605368`.

[49] Ole J. Mengshoel. Understanding the Role of Noise in Stochastic Local Search: Analysis and Experiments. *Artificial Intelligence*, 172(8):955–990, 2008. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2007.09.010. URL `https://www.sciencedirect.com/science/article/pii/S0004370208000040`.

[50] Ole Jakob Mengshoel, Eirik Flogard, Jon Riege, and Tong Yu. Stochastic Local Search Heuristics for Efficient Feature Selection: An Experimental Study. In *NIK Norsk informatikkonferanse*, 2021.

[51] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A Dialog Research Software Platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2014. URL `https://aclanthology.org/D17-2014`.

[52] Council of Europe. The CEFR Levels. `https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions`, 2022. Accessed: 2022-05-18.

[53] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[54] Y.-S. Park and S. Lek. Chapter 7 - Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. In *Ecological Model Types*, volume 28 of *Developments in Environmental Modelling*, pages 123–140. Elsevier, 2016. doi: https://doi.org/10.1016/B978-0-444-63623-2.00007-4. URL `https://www.sciencedirect.com/science/article/pii/B9780444636232000074`.

[55] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. SOLOIST: Few-shot Task-Oriented Dialog with a Single Pre-trained Auto-regressive Model. *CoRR*, abs/2005.05298, 2020. URL `https://arxiv.org/abs/2005.05298`.

[56] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

[57] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020.

[58] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.39. URL `https://aclanthology.org/2021.eacl-main.39`.

[59] Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL `https://aclanthology.org/P19-1493`.

[60] Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. What to Pre-Train on? Efficient Intermediate Task Selection. In *Proceedings of*

the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10585–10605, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.emnlp-main.827`.

[61] Alec Radford and Karthik Narasimhan. Improving Language Understanding by Generative Pre-Training, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

[62] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

[63] Rehan Sadiq, Manuel J. Rodriguez, and Haroon R. Mian. Empirical Models to Predict Disinfection By-Products (DBPs) in Drinking Water: An Updated Review. In *Encyclopedia of Environmental Health (Second Edition)*, pages 324–338. Elsevier, Oxford, second edition edition, 2019. ISBN 978-0-444-63952-3. doi: https://doi.org/10.1016/B978-0-12-409548-9.11193-5. URL `https://www.sciencedirect.com/science/article/pii/B9780124095489111935`.

[64] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700. Springer Nature, 2019.

[65] Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention, 2020. URL `https://proceedings.neurips.cc/paper/2020/file/460191c72f67e90150a093b4585e7eb4-Paper.pdf`.

[66] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*, pages 194–206, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32381-3.

[67] Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. Adding Chit-Chat to Enhance Task-Oriented Dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.124. URL `https://aclanthology.org/2021.naacl-main.124`.

[68] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf`.

[69] Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.129. URL `https://aclanthology.org/2020.findings-emnlp.129`.

[70] Yen-Jen Tai and Hung-Yu Kao. Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation. In *Proceedings of International Conference on Information Integration and Web-Based Applications & Services*, IIWAS '13, page 53–62, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321136. doi: 10.1145/2539150.2539190. URL `https://doi.org/10.1145/2539150.2539190`.

[71] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language Models for Dialog Applications. *CoRR*, abs/2201.08239, 2022. URL `https://arxiv.org/abs/2201.08239`.

[72] Sebastian Thrun and Lorien Pratt. *Learning to Learn.* Kluwer Academic Publishers, USA, 1998. ISBN 0792380479.

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in*

Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[74] Steven Walczak and Narciso Cerpa. Artificial Neural Networks. In *Encyclopedia of Physical Science and Technology (Third Edition)*, pages 631–645. Academic Press, New York, 2003. ISBN 978-0-12-227410-7. doi: https://doi.org/10.1016/B0-12-227410-5/00837-1. URL https://www.sciencedirect.com/science/article/pii/B0122274105008371.

[75] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *CoRR*, abs/1901.08149, 2019. URL http://arxiv.org/abs/1901.08149.

[76] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[77] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3506–3510, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025496. URL https://doi.org/10.1145/3025453.3025496.

[78] Weilai Xu, Charlie Hargood, Wen Tang, and Fred Charles. Towards Generating Stylistic Dialogues for Narratives Using Data-Driven Approaches. In *International Conference on Interactive Storytelling*, 2018. URL http://eprints.bournemouth.ac.uk/31506/1/icids-paper-66.pdf.

[79] Weilai Xu, Fred Charles, Charlie Hargood, Feng Tian, and Wen Tang. Influence of Personality-Based Features for Dialogue Generation in Computational Narratives. In *European Conference on Artificial Intelligence*, 2020. URL https://ecai2020.eu/papers/1274_paper.pdf.

[80] Sina Zarrieß, Henrik Voigt, and Simeon Schüz. Decoding Methods in Neural Language Generation: A Survey. *Information*, 12(9), 2021. ISSN 2078-2489.

doi: 10.3390/info12090355. URL https://www.mdpi.com/2078-2489/12/9/355.

[81] Boliang Zhang, Ying Lyu, Ning Ding, Tianhao Shen, Zhaoyang Jia, Kun Han, and Kevin Knight. A Hybrid Task-Oriented Dialog System with Domain and Task Adaptive Pretraining. *CoRR*, abs/2102.04506, 2021. URL https://arxiv.org/abs/2102.04506.

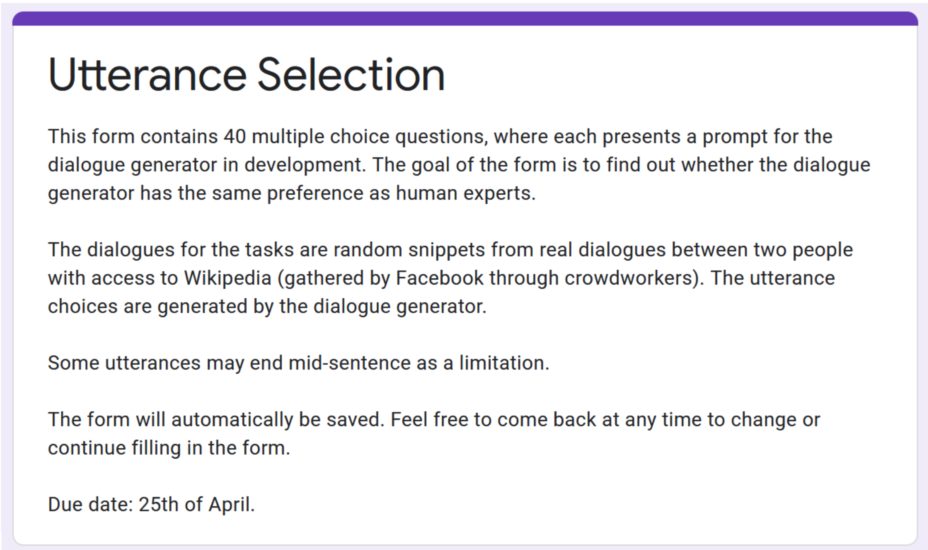
[82] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[83] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Appendices

A.1 Experiment 2: Questionnaire

This appendix contains the questionnaire conducted as part of Experiment 2 (see Section 5.3). The participants of the questionnaire are given the instructions in Figure 1 before they can solve the forty tasks. All of the tasks follow the same structure as the task seen in Figure 2. The only difference between the tasks is the content (i.e., the topic, CEFR level, dialogue snippet, and utterance options shown).



Utterance Selection

This form contains 40 multiple choice questions, where each presents a prompt for the dialogue generator in development. The goal of the form is to find out whether the dialogue generator has the same preference as human experts.

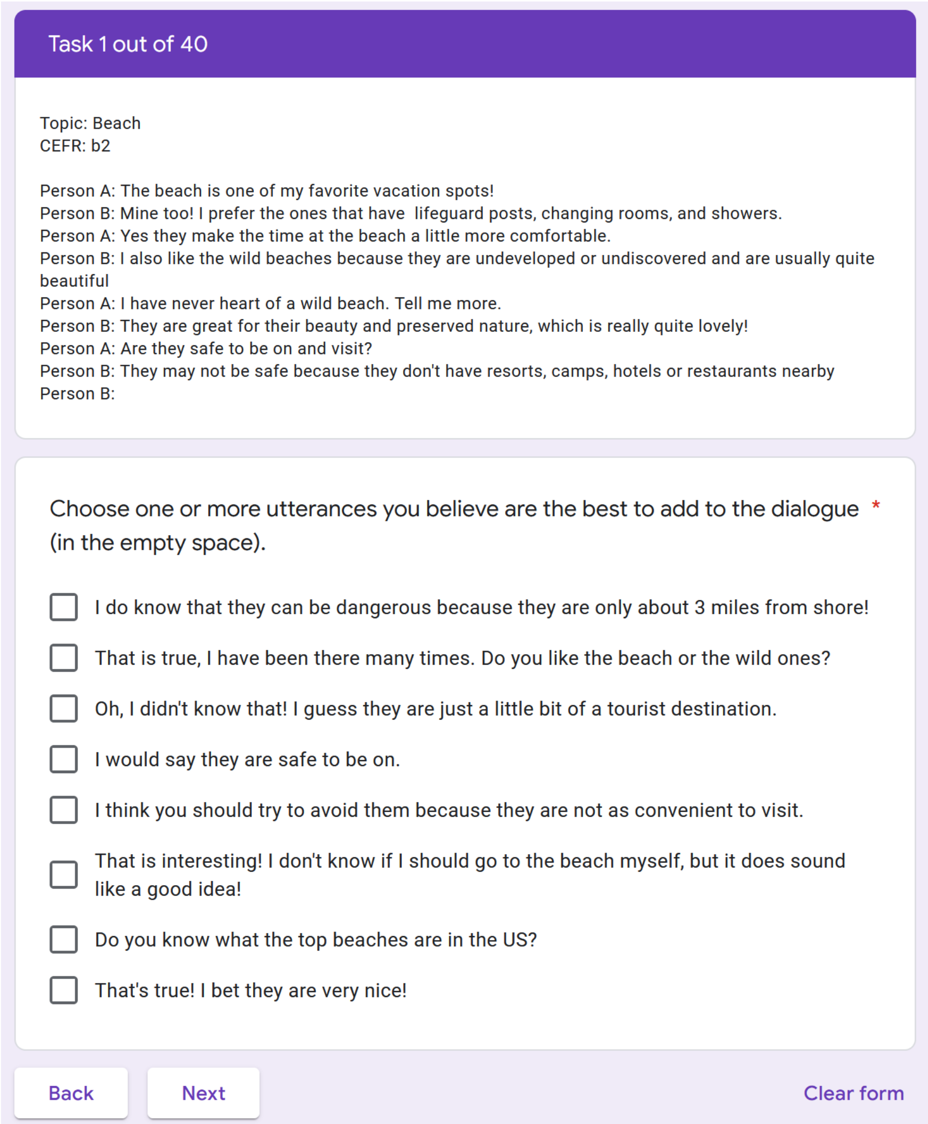
The dialogues for the tasks are random snippets from real dialogues between two people with access to Wikipedia (gathered by Facebook through crowdworkers). The utterance choices are generated by the dialogue generator.

Some utterances may end mid-sentence as a limitation.

The form will automatically be saved. Feel free to come back at any time to change or continue filling in the form.

Due date: 25th of April.

Figure 1: The instructions given to the participants of the questionnaire of Experiment 2.



Task 1 out of 40

Topic: Beach
CFR: b2

Person A: The beach is one of my favorite vacation spots!
Person B: Mine too! I prefer the ones that have lifeguard posts, changing rooms, and showers.
Person A: Yes they make the time at the beach a little more comfortable.
Person B: I also like the wild beaches because they are undeveloped or undiscovered and are usually quite beautiful
Person A: I have never heart of a wild beach. Tell me more.
Person B: They are great for their beauty and preserved nature, which is really quite lovely!
Person A: Are they safe to be on and visit?
Person B: They may not be safe because they don't have resorts, camps, hotels or restaurants nearby
Person B:

Choose one or more utterances you believe are the best to add to the dialogue *
(in the empty space).

- [ ] I do know that they can be dangerous because they are only about 3 miles from shore!
- [ ] That is true, I have been there many times. Do you like the beach or the wild ones?
- [ ] Oh, I didn't know that! I guess they are just a little bit of a tourist destination.
- [ ] I would say they are safe to be on.
- [ ] I think you should try to avoid them because they are not as convenient to visit.
- [ ] That is interesting! I don't know if I should go to the beach myself, but it does sound like a good idea!
- [ ] Do you know what the top beaches are in the US?
- [ ] That's true! I bet they are very nice!

Back Next Clear form

Figure 2: The first task of the questionnaire of Experiment 2. The participants are able to select one or more utterances they believe are the best to add to the dialogue snippet (on the top), with its topic and CEFR level.

Task 2 out of 40

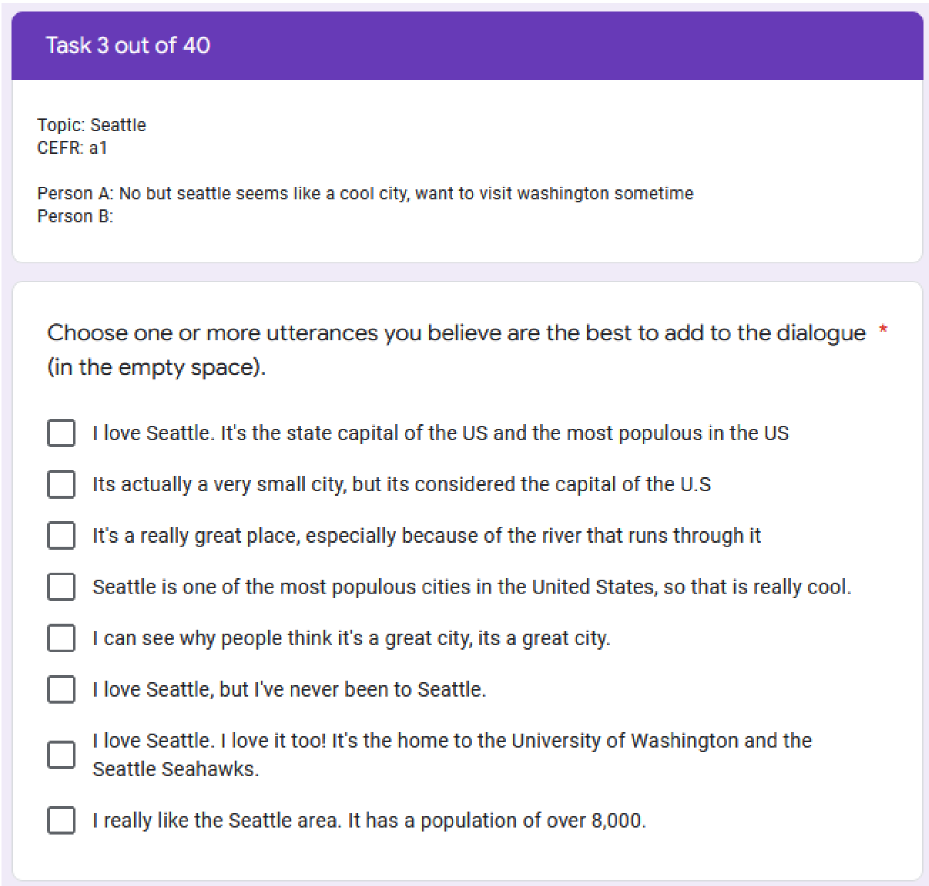
Topic: Beach
CEFR: b2

Person A:
Person B: The beach is one of my favorite vacation spots!
Person A: Mine too! I prefer the ones that have lifeguard posts, changing rooms, and showers.
Person B: Yes they make the time at the beach a little more comfortable.
Person A: I also like the wild beaches because they are undeveloped or undiscovered and are usually quite beautiful
Person B: I have never heart of a wild beach. Tell me more.
Person A: They are great for their beauty and preserved nature, which is really quite lovely!
Person B: Are they safe to be on and visit?
Person A: They may not be safe because they don't have resorts, camps, hotels or restaurants nearby

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] I love to go to the beach! It's one of the largest and most popular tourist destinations in the world.
- [ ] I have never been to the beach, but I like the view from the beach. Do you?
- [ ] I love beaches! I love the beach with its natural beauty and waterfalls!
- [ ] I love to go to the beach at night. The water is beautiful and the ocean is beautiful.
- [ ] I really love the beaches on the beach. They are a good place to stay for your vacation!
- [ ] I love the beach and the beach is the most popular tourist attraction in the world.
- [ ] I love the beach, especially the sandy beaches, and the waves.
- [ ] I love beaches, especially the beaches of Hawaii.

Figure 3: The second task of the questionnaire of Experiment 2.



Task 3 out of 40

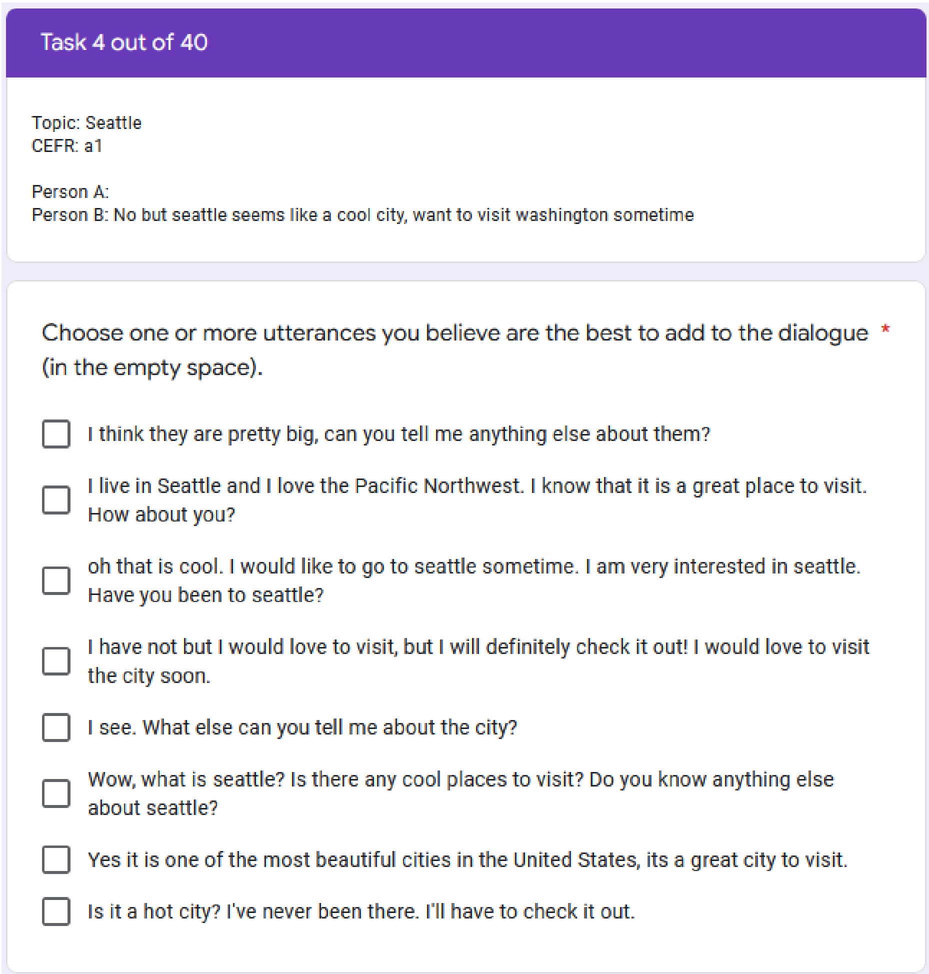
Topic: Seattle
CEFR: a1

Person A: No but seattle seems like a cool city, want to visit washington sometime
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I love Seattle. It's the state capital of the US and the most populous in the US
- [ ] Its actually a very small city, but its considered the capital of the U.S
- [ ] It's a really great place, especially because of the river that runs through it
- [ ] Seattle is one of the most populous cities in the United States, so that is really cool.
- [ ] I can see why people think it's a great city, its a great city.
- [ ] I love Seattle, but I've never been to Seattle.
- [ ] I love Seattle. I love it too! It's the home to the University of Washington and the Seattle Seahawks.
- [ ] I really like the Seattle area. It has a population of over 8,000.

Figure 4: The third task of the questionnaire of Experiment 2.



Task 4 out of 40

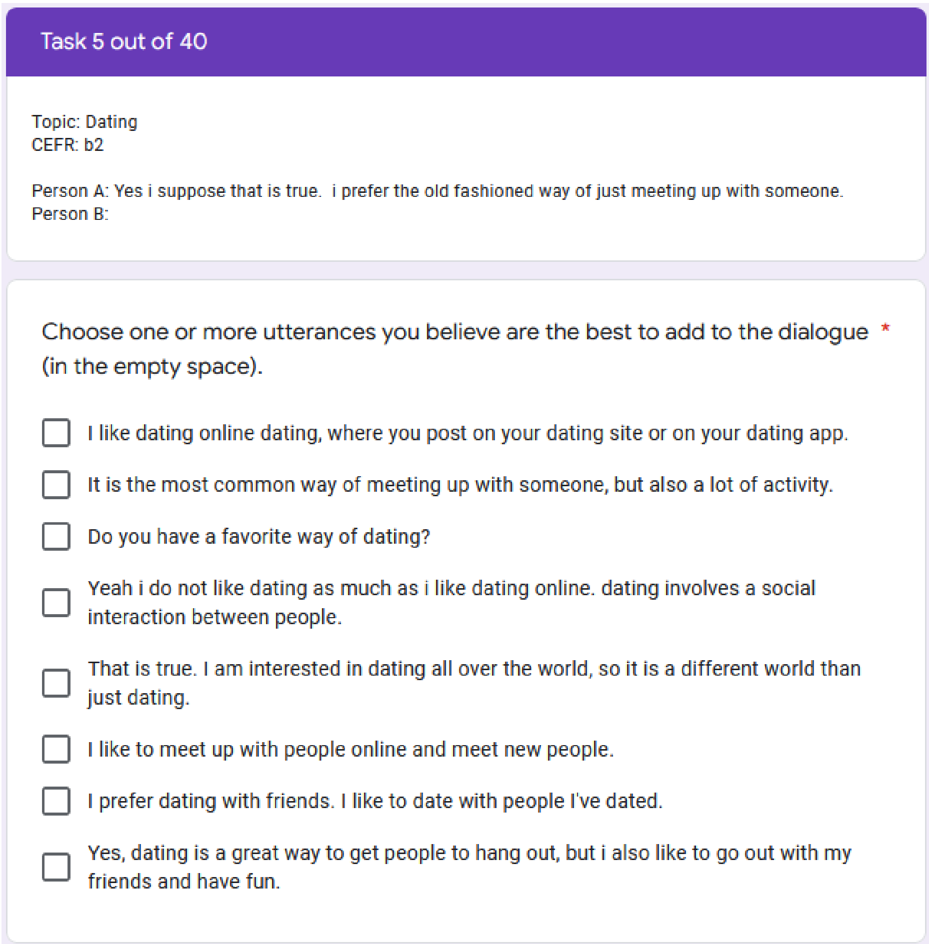
Topic: Seattle
CEFR: a1

Person A:
Person B: No but seattle seems like a cool city, want to visit washington sometime

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I think they are pretty big, can you tell me anything else about them?
- [ ] I live in Seattle and I love the Pacific Northwest. I know that it is a great place to visit. How about you?
- [ ] oh that is cool. I would like to go to seattle sometime. I am very interested in seattle. Have you been to seattle?
- [ ] I have not but I would love to visit, but I will definitely check it out! I would love to visit the city soon.
- [ ] I see. What else can you tell me about the city?
- [ ] Wow, what is seattle? Is there any cool places to visit? Do you know anything else about seattle?
- [ ] Yes it is one of the most beautiful cities in the United States, its a great city to visit.
- [ ] Is it a hot city? I've never been there. I'll have to check it out.

Figure 5: The fourth task of the questionnaire of Experiment 2.



Task 5 out of 40

Topic: Dating
CEFR: b2

Person A: Yes i suppose that is true. i prefer the old fashioned way of just meeting up with someone.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] I like dating online dating, where you post on your dating site or on your dating app.
- [ ] It is the most common way of meeting up with someone, but also a lot of activity.
- [ ] Do you have a favorite way of dating?
- [ ] Yeah i do not like dating as much as i like dating online. dating involves a social interaction between people.
- [ ] That is true. I am interested in dating all over the world, so it is a different world than just dating.
- [ ] I like to meet up with people online and meet new people.
- [ ] I prefer dating with friends. I like to date with people I've dated.
- [ ] Yes, dating is a great way to get people to hang out, but i also like to go out with my friends and have fun.

Figure 6: The fifth task of the questionnaire of Experiment 2.

Task 6 out of 40

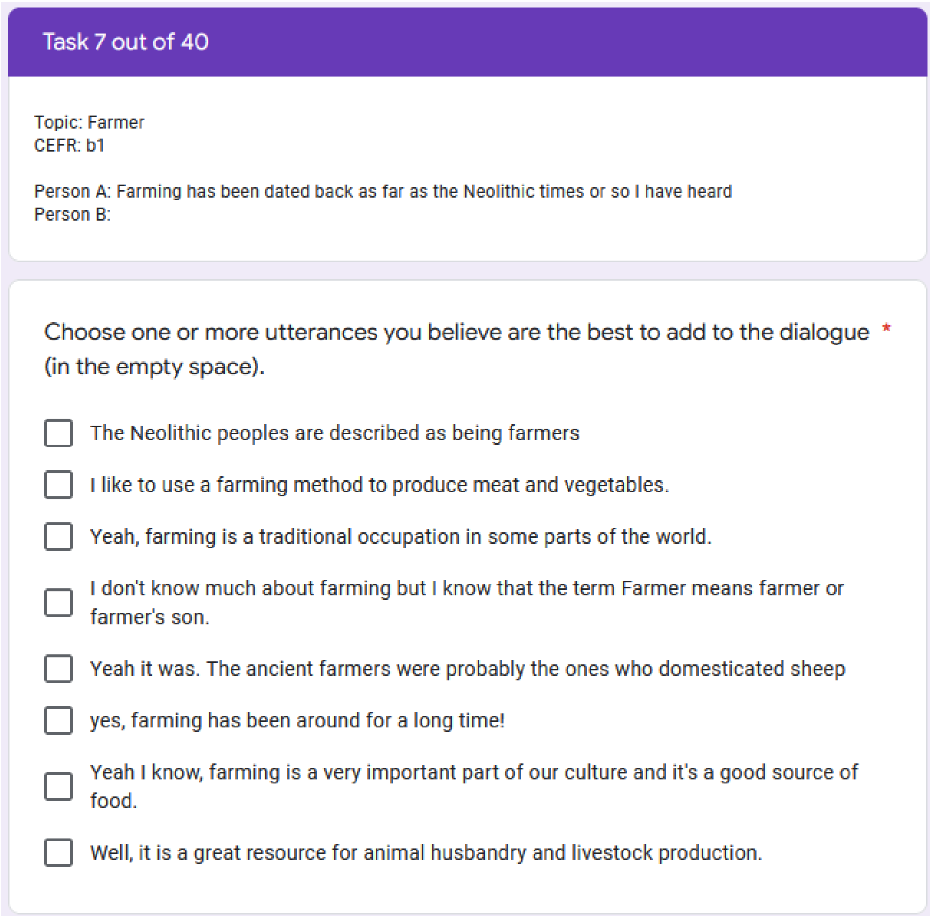
Topic: Dating
CEFR: b2

Person A:
Person B: Yes i suppose that is true. i prefer the old fashioned way of just meeting up with someone.

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Do you like to date other people? I am not sure if i know which one i like. I know that dating is not always easy. I love to get my dates dates done. i like to meet new people.
- [ ] I guess it is. Do you prefer the traditional dating method or the modern dating system? I'm not sure.
- [ ] Yes that is true. I like to meet new people and try to get as close to them as possible. It's also why I am so comfortable with a new person. I love meeting new people. If you go to a
- [ ] I am not sure, but I do know that dating is a very popular form of social interaction. I wish I could find my way around the world!
- [ ] I have met people online and it is a great way to get to know one another. It seems like it is a great way to get to know one another. Do you have a favorite way to meet?
- [ ] That is true, but I think it is better to do it at home. I would rather go to a restaurant and have a drink than go to a bar and go to the pub. I would prefer a place where I can meet up
- [ ] I think it is hard to find the right person to date. You just need to be a good person. I'm not sure that's the case with my boyfriend. I love his face though.
- [ ] Yea i bet it's a lot easier to date with someone. i can go to the pool and watch a movie and have a drink with my boyfriend

Figure 7: The sixth task of the questionnaire of Experiment 2.



Task 7 out of 40

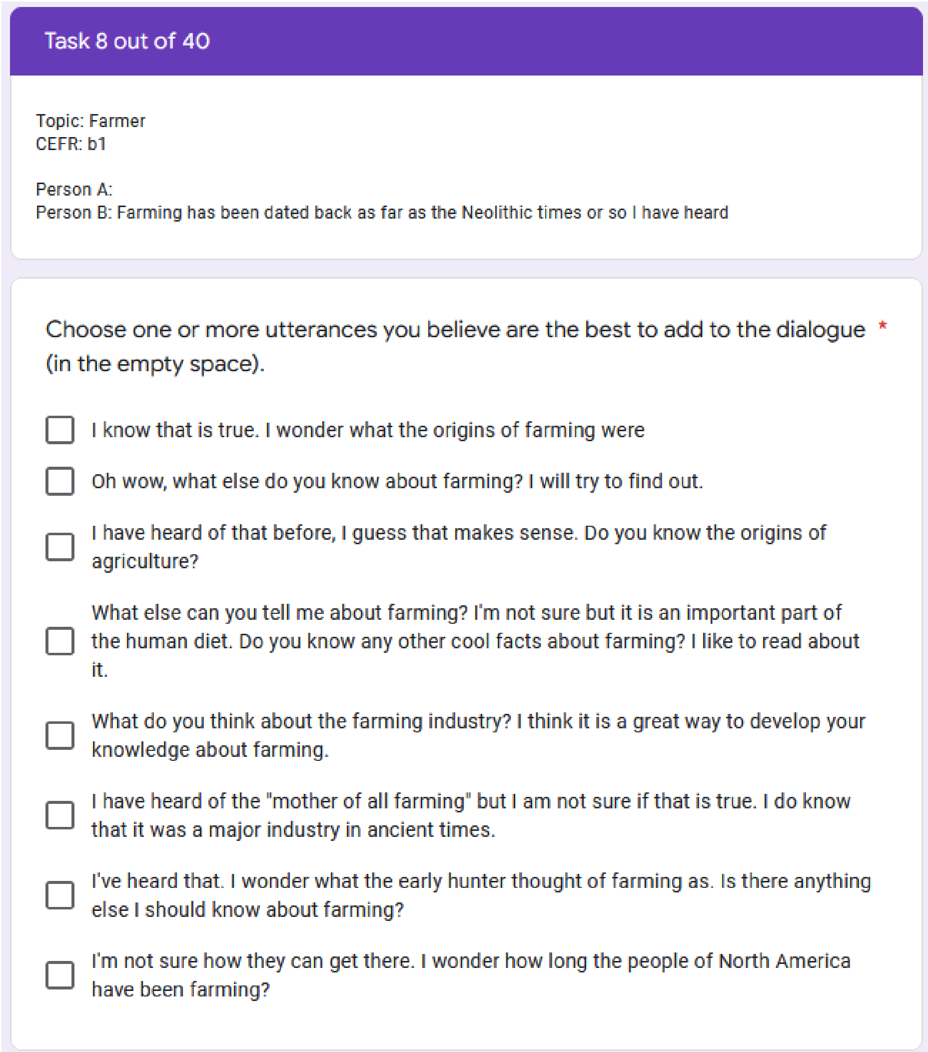
Topic: Farmer
CEFR: b1

Person A: Farming has been dated back as far as the Neolithic times or so I have heard
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] The Neolithic peoples are described as being farmers
- [ ] I like to use a farming method to produce meat and vegetables.
- [ ] Yeah, farming is a traditional occupation in some parts of the world.
- [ ] I don't know much about farming but I know that the term Farmer means farmer or farmer's son.
- [ ] Yeah it was. The ancient farmers were probably the ones who domesticated sheep
- [ ] yes, farming has been around for a long time!
- [ ] Yeah I know, farming is a very important part of our culture and it's a good source of food.
- [ ] Well, it is a great resource for animal husbandry and livestock production.

Figure 8: The seventh task of the questionnaire of Experiment 2.



Task 8 out of 40

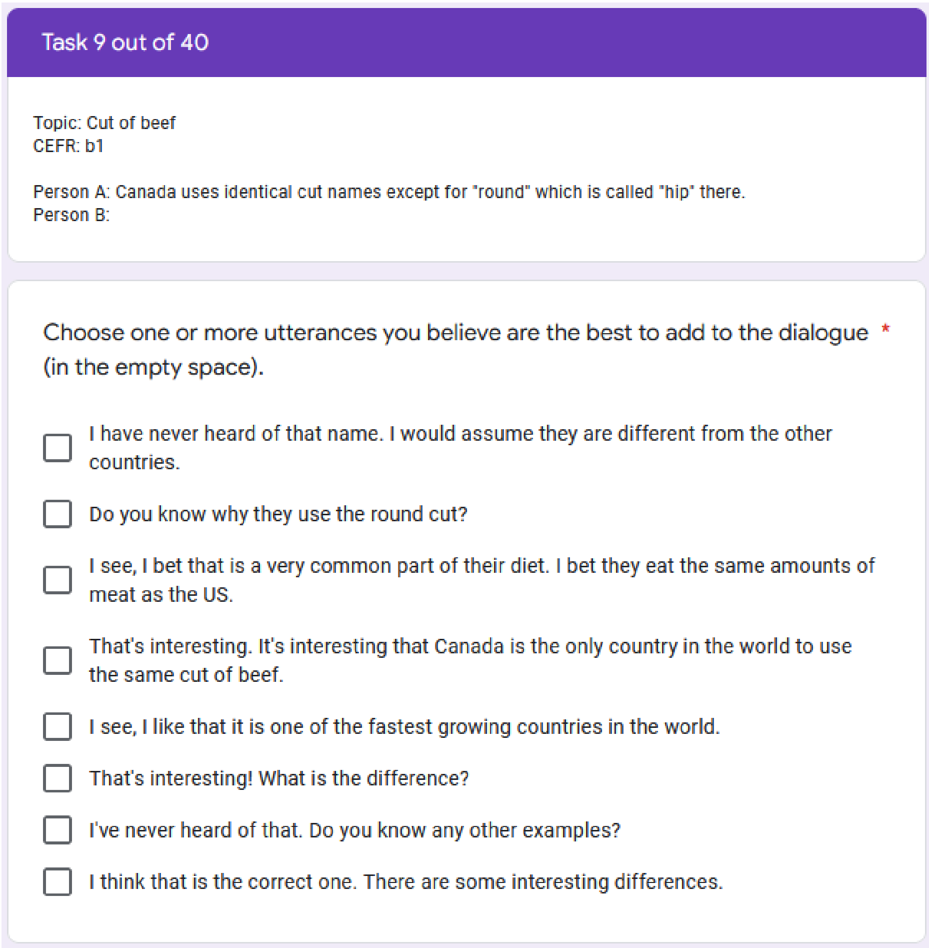
Topic: Farmer
CEFR: b1

Person A:
Person B: Farming has been dated back as far as the Neolithic times or so I have heard

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I know that is true. I wonder what the origins of farming were
- [ ] Oh wow, what else do you know about farming? I will try to find out.
- [ ] I have heard of that before, I guess that makes sense. Do you know the origins of agriculture?
- [ ] What else can you tell me about farming? I'm not sure but it is an important part of the human diet. Do you know any other cool facts about farming? I like to read about it.
- [ ] What do you think about the farming industry? I think it is a great way to develop your knowledge about farming.
- [ ] I have heard of the "mother of all farming" but I am not sure if that is true. I do know that it was a major industry in ancient times.
- [ ] I've heard that. I wonder what the early hunter thought of farming as. Is there anything else I should know about farming?
- [ ] I'm not sure how they can get there. I wonder how long the people of North America have been farming?

Figure 9: The eighth task of the questionnaire of Experiment 2.



Task 9 out of 40

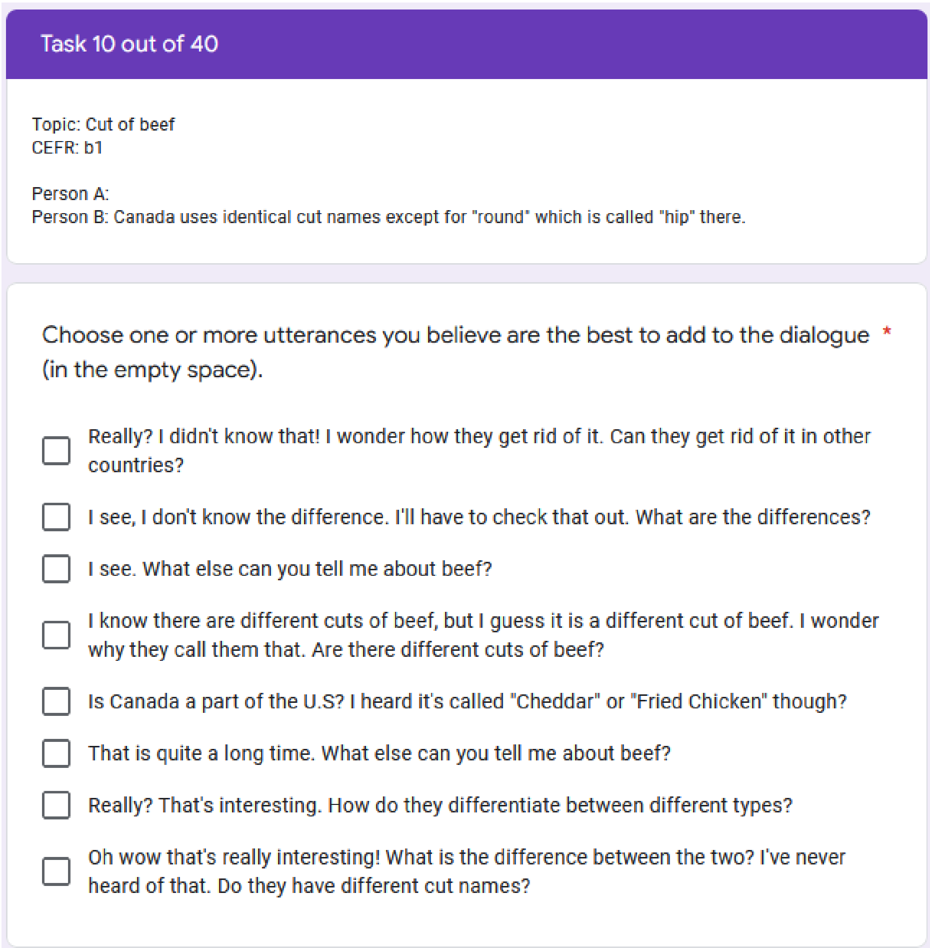
Topic: Cut of beef
CEFR: b1

Person A: Canada uses identical cut names except for "round" which is called "hip" there.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I have never heard of that name. I would assume they are different from the other countries.
- [ ] Do you know why they use the round cut?
- [ ] I see, I bet that is a very common part of their diet. I bet they eat the same amounts of meat as the US.
- [ ] That's interesting. It's interesting that Canada is the only country in the world to use the same cut of beef.
- [ ] I see, I like that it is one of the fastest growing countries in the world.
- [ ] That's interesting! What is the difference?
- [ ] I've never heard of that. Do you know any other examples?
- [ ] I think that is the correct one. There are some interesting differences.

Figure 10: The ninth task of the questionnaire of Experiment 2.



Task 10 out of 40

Topic: Cut of beef
CEFR: b1

Person A:
Person B: Canada uses identical cut names except for "round" which is called "hip" there.

Choose one or more utterances you believe are the best to add to the dialogue *
(in the empty space).

- [ ] Really? I didn't know that! I wonder how they get rid of it. Can they get rid of it in other countries?
- [ ] I see, I don't know the difference. I'll have to check that out. What are the differences?
- [ ] I see. What else can you tell me about beef?
- [ ] I know there are different cuts of beef, but I guess it is a different cut of beef. I wonder why they call them that. Are there different cuts of beef?
- [ ] Is Canada a part of the U.S? I heard it's called "Cheddar" or "Fried Chicken" though?
- [ ] That is quite a long time. What else can you tell me about beef?
- [ ] Really? That's interesting. How do they differentiate between different types?
- [ ] Oh wow that's really interesting! What is the difference between the two? I've never heard of that. Do they have different cut names?

Figure 11: The 10th task of the questionnaire of Experiment 2.

Task 11 out of 40

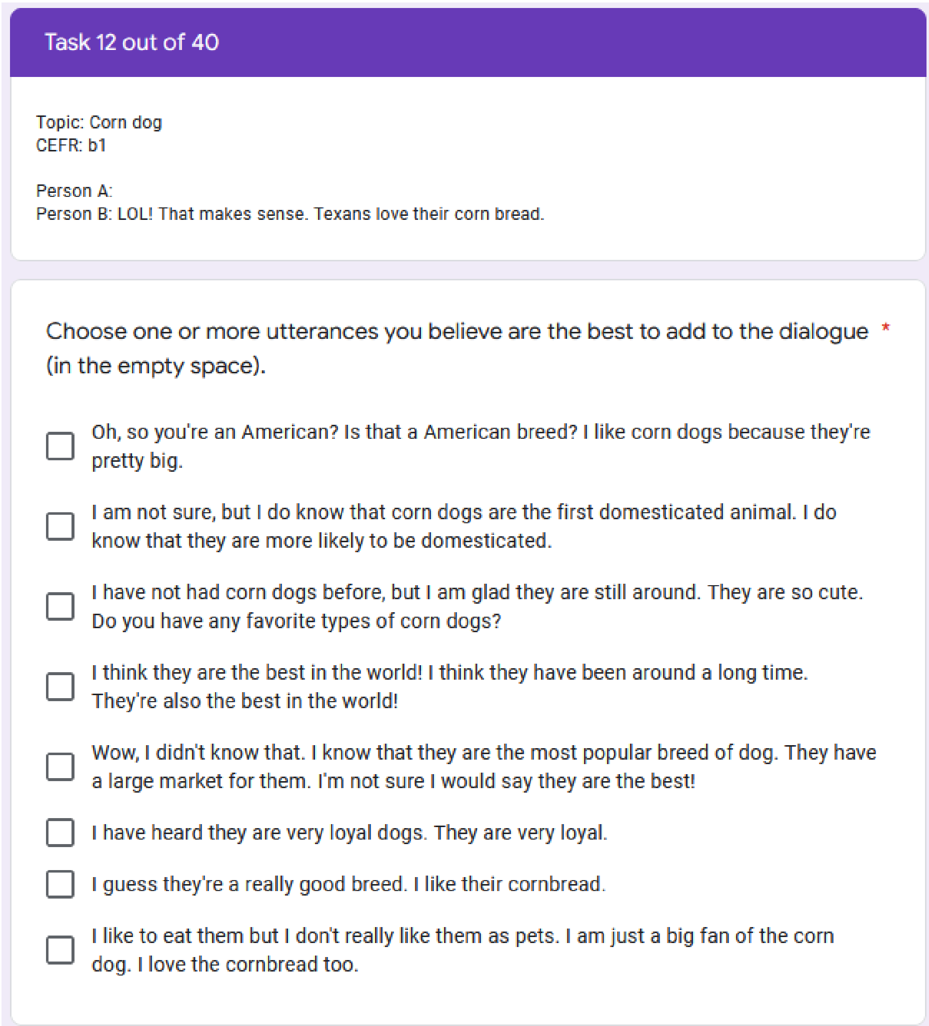
Topic: Corn dog
CFFR: b1

Person A: LOL! That makes sense. Texans love their corn bread.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes, I agree. It is very popular and used as a food source in many parts of the world.
- [ ] Yes, corn dogs are very good at keeping out pests, like mosquitoes, and bugs.
- [ ] I think the corn dog is the most popular breed of dog!
- [ ] The Corn Dog is a small dog, so it's a good dog to have around!
- [ ] Yes, and their sweet corn bread is so good. Do you know what kind of corn bread they use?
- [ ] I love corn bread too! The corn dog is a companion dog for the corn husk.
- [ ] I've never heard of that, but I have heard it is the most common and most widely consumed food in the US
- [ ] I love corn bread too! I have heard that corn dogs are actually called "chicken dogs" in the US.

Figure 12: The 11th task of the questionnaire of Experiment 2.



Task 12 out of 40

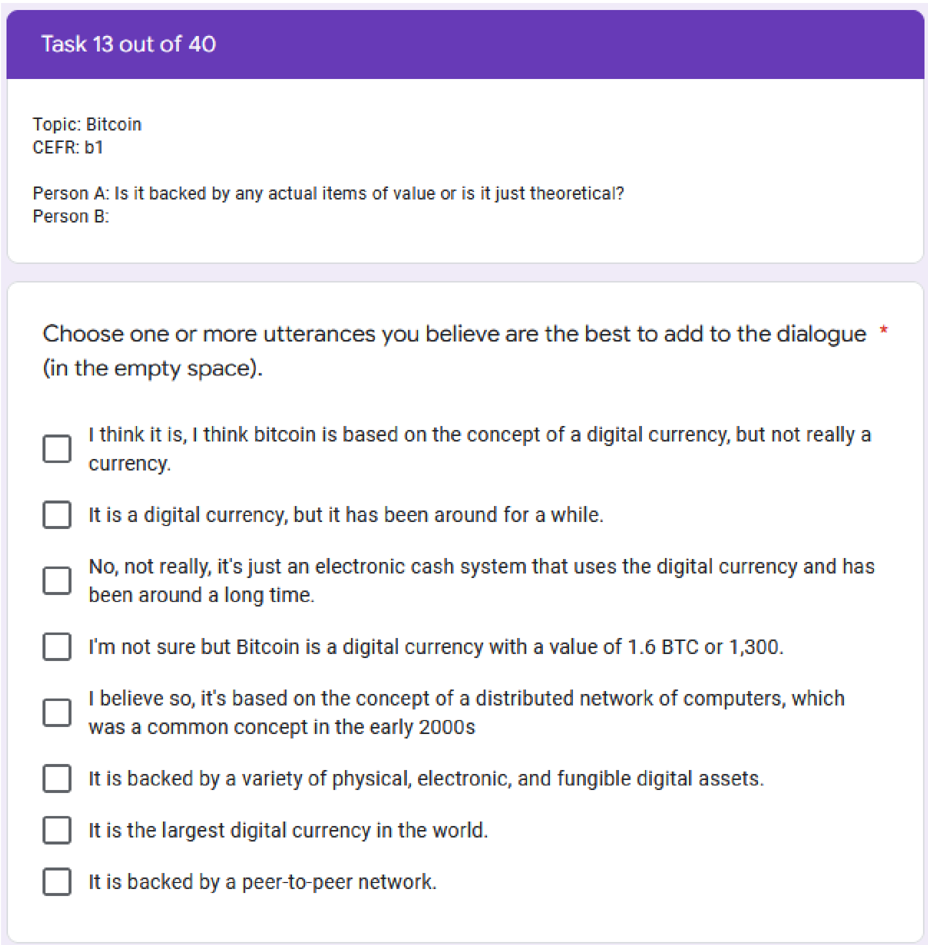
Topic: Corn dog
CEFR: b1

Person A:
Person B: LOL! That makes sense. Texans love their corn bread.

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] Oh, so you're an American? Is that a American breed? I like corn dogs because they're pretty big.
- [ ] I am not sure, but I do know that corn dogs are the first domesticated animal. I do know that they are more likely to be domesticated.
- [ ] I have not had corn dogs before, but I am glad they are still around. They are so cute. Do you have any favorite types of corn dogs?
- [ ] I think they are the best in the world! I think they have been around a long time. They're also the best in the world!
- [ ] Wow, I didn't know that. I know that they are the most popular breed of dog. They have a large market for them. I'm not sure I would say they are the best!
- [ ] I have heard they are very loyal dogs. They are very loyal.
- [ ] I guess they're a really good breed. I like their cornbread.
- [ ] I like to eat them but I don't really like them as pets. I am just a big fan of the corn dog. I love the cornbread too.

Figure 13: The 12th task of the questionnaire of Experiment 2.



Task 13 out of 40

Topic: Bitcoin
CEFR: b1

Person A: Is it backed by any actual items of value or is it just theoretical?
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I think it is, I think bitcoin is based on the concept of a digital currency, but not really a currency.
- [ ] It is a digital currency, but it has been around for a while.
- [ ] No, not really, it's just an electronic cash system that uses the digital currency and has been around a long time.
- [ ] I'm not sure but Bitcoin is a digital currency with a value of 1.6 BTC or 1,300.
- [ ] I believe so, it's based on the concept of a distributed network of computers, which was a common concept in the early 2000s
- [ ] It is backed by a variety of physical, electronic, and fungible digital assets.
- [ ] It is the largest digital currency in the world.
- [ ] It is backed by a peer-to-peer network.

Figure 14: The 13th task of the questionnaire of Experiment 2.

Task 14 out of 40

Topic: Bitcoin
CEFR: b1

Person A:
Person B: Is it backed by any actual items of value or is it just theoretical?

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] It has a lot of value and it's just a speculative way of looking at things.
- [ ] I don't think it would be hard to prove that, but it is quite a complicated process.
- [ ] I think so, but it is only being mined on a computer for profit.
- [ ] I think it is a pretty cool idea. Bitcoin is an interesting coin.
- [ ] I'm not sure but I think the price has been rising for a while. It seems like the price has increased.
- [ ] it is a digital currency and is based on the digital currency bitcoin
- [ ] Well, they are all coins. There are a lot of exchanges that accept bitcoin, and the majority of them are closed, so I'm not sure if there are any. There are also many bitcoin-related exchanges.
- [ ] Yes, I agree. It is very efficient and easy to use.

Figure 15: The 14th task of the questionnaire of Experiment 2.

Task 15 out of 40

Topic: Pit bull
CEFR: b1

Person A: I love them! They get so much hate but its all bad owners that make them aggressive.
Person B: Yes, they're giant sweeties. And a lot of dogs are considered to be pit bulls such as pit bull terriers, American bullys and staffordshire terriers
Person A: Really? Pit bulls look a lot different to terriers and American bulldogs.
Person B: Yes it's totally strange, but even American bulldogs are sometimes included as pit bulls.
Person A:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes, I love American bulldogs too. They are such a cute breed and are often used in the field.
- [ ] Yeah, they are a bit of a family dog breed. They are a bit more prominent in the US and Canada, but still have their own family.
- [ ] Yes I see. I'm sure that some people find it sad that they have to go to such lengths to get their dogs to eat, but they do have a great sense of humor.
- [ ] I don't think it would be so strange to see a pit bull as a pit bull.
- [ ] I was thinking of adding a pit bull to my family. They are my favorite breed.
- [ ] I have seen that a lot of pit bulls are called "pit bull terriers" as well. I really love pit bulls and it's great to have them around.
- [ ] I guess they are because they are so cute. I like to think they are part of the family.
- [ ] I think pit bulls are the best breed for their age. I am a pit bull owner myself.

Figure 16: The 15th task of the questionnaire of Experiment 2.

Task 16 out of 40

Topic: Pit bull
CEFR: b1

Person A:
Person B: I love them! They get so much hate but its all bad owners that make them aggressive.
Person A: Yes, they're giant sweeties. And a lot of dogs are considered to be pit bulls such as pit bull terriers, American bullys and staffordshire terriers
Person B: Really? Pit bulls look a lot different to terriers and American bulldogs.
Person A: Yes it's totally strange, but even American bulldogs are sometimes included as pit bulls.

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I love pit bulls, they are so cute. Pit bulls are so cute. They are the most popular breed of pit bull.
- [ ] They are! I think pit bulls are the most aggressive breed of dog and the most loyal to their owner. They are so beautiful!
- [ ] They are also known for their aggressive behavior. They are considered the worst breed of dog.
- [ ] I have a pit bull and a terrier. I had one when I was a kid. I've always loved dogs and I had a pit bull and a terrier. I've always wanted to have one.
- [ ] Yes, I am a pit bull owner, but I don't know much about them. They are so sweet and aggressive.
- [ ] I'm not really a dog owner. I know that pit bulls are very aggressive, but it's hard to imagine they'd be a pit bull!
- [ ] I have a pit bull. It's one of my favorite dogs. They have a big heart.
- [ ] Pit bulls are very nice dogs. They are very loyal and they have been bred for obedience and obedience to humans.

Figure 17: The 16th task of the questionnaire of Experiment 2.

Task 17 out of 40

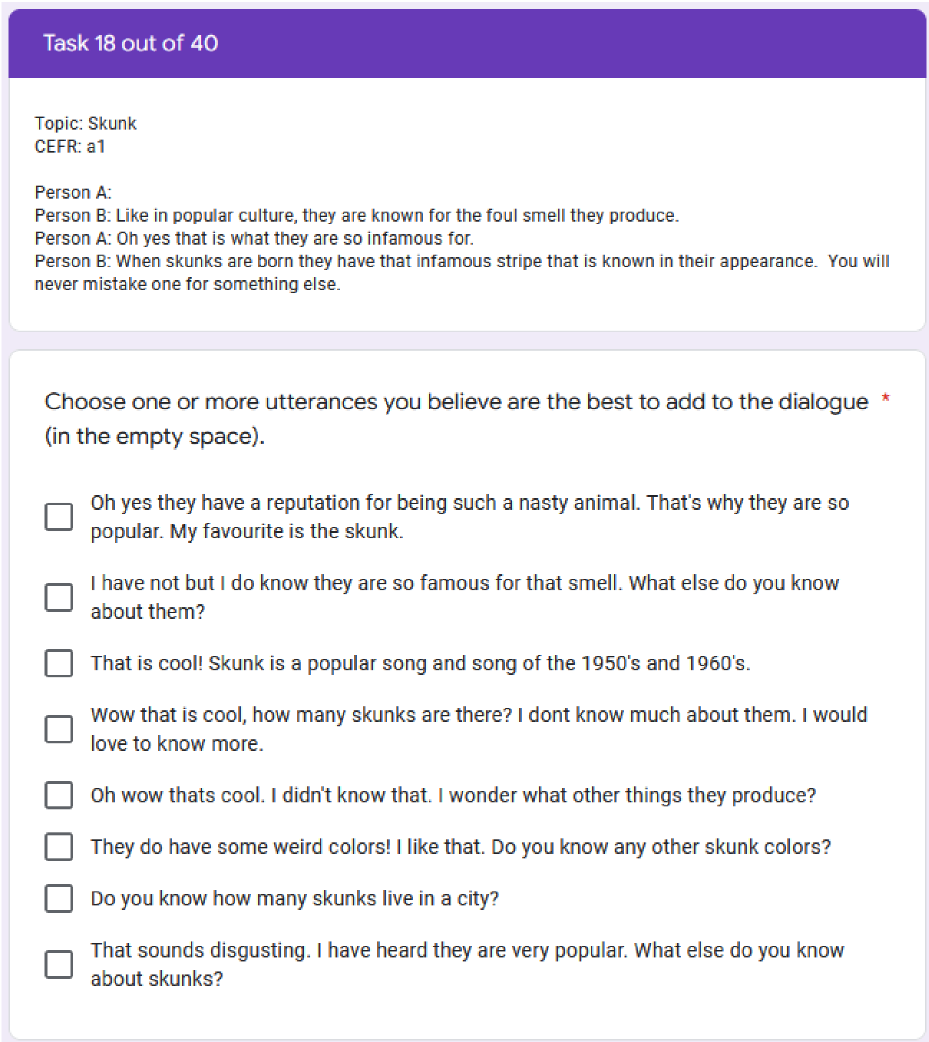
Topic: Skunk
CEFR: a1

Person A: Like in popular culture, they are known for the foul smell they produce.
Person B: Oh yes that is what they are so infamous for.
Person A: When skunks are born they have that infamous stripe that is known in their appearance. You will never mistake one for something else.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] Skunks are a unique species of lizard. They are usually found in tropical and temperate climates.
- [ ] I've heard that they can be very aggressive in their behavior.
- [ ] Skunks have a thick, furry, long tail and an unruly tail. They are known to have been bred for hunting and for war.
- [ ] The difference between skunks and the other animals is usually just the color of the skin.
- [ ] Skunks are considered to be the most intelligent of mammals.
- [ ] Skunks have a long life expectancy. They are known for being able to survive in colder climates.
- [ ] Skunks are also known for being a large and powerful creature.
- [ ] I think that's a good thing. Skunks are often the most widely consumed species of fish.

Figure 18: The 17th task of the questionnaire of Experiment 2.



Task 18 out of 40

Topic: Skunk
CEFR: a1

Person A:
Person B: Like in popular culture, they are known for the foul smell they produce.
Person A: Oh yes that is what they are so infamous for.
Person B: When skunks are born they have that infamous stripe that is known in their appearance. You will never mistake one for something else.

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Oh yes they have a reputation for being such a nasty animal. That's why they are so popular. My favourite is the skunk.
- [ ] I have not but I do know they are so famous for that smell. What else do you know about them?
- [ ] That is cool! Skunk is a popular song and song of the 1950's and 1960's.
- [ ] Wow that is cool, how many skunks are there? I dont know much about them. I would love to know more.
- [ ] Oh wow thats cool. I didn't know that. I wonder what other things they produce?
- [ ] They do have some weird colors! I like that. Do you know any other skunk colors?
- [ ] Do you know how many skunks live in a city?
- [ ] That sounds disgusting. I have heard they are very popular. What else do you know about skunks?

Figure 19: The 18th task of the questionnaire of Experiment 2.

Task 19 out of 40

Topic: Ford Mustang
CEFR: a2

Person A: Who came up with the name "Mustang"? It's an odd name for a car!
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I'm not sure but the Mustang was a brand name for a brand of motorized muscle car.
- [ ] It is a sport utility vehicle. It's a sedan. It's been around for over 100 years.
- [ ] I'm not sure but it's a Ford Mustang, which is a Type-A, but the name comes from the fact that the Mustang is a modified version of the Mustang.
- [ ] I'm not sure but I know it's actually a classic American sports car produced by Ford in 1963.
- [ ] It was a mid-size car, but it was a great family car! It was a classic American car, and I was lucky enough to have it in my home!
- [ ] Ford Mustang is a compact car.
- [ ] It was actually invented by Ford in 1962!
- [ ] Ford Mustang is a Japanese automaker.

Figure 20: The 19th task of the questionnaire of Experiment 2.

Task 20 out of 40

Topic: Ford Mustang
CEFR: a2

Person A:
Person B: Who came up with the name "Mustang"? It's an odd name for a car!

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] Yes it was actually a Ford Mustang, I remember when they used the "Mustang" logo on the front bumper. But I do remember the engine being the same as the one on the back of the car. They also have the "
- [ ] Interesting! I didn't know that. I was so surprised! What is the origin of the name "Mustang"? I would have thought it was just a reference to the Mustang. It was produced in 1969.
- [ ] Wow that's really cool. I think it's a very well-known brand. I wonder what other brands it has?
- [ ] That's pretty interesting. I'm sure they're pretty popular. I wonder how popular they are. I guess they get some money. I wonder how popular they are. I like the old Ford Mustang. Is it a
- [ ] I think the Ford Mustang is still the best car of the 1980s. My favorite car was the Ford Mustang.
- [ ] I'm not sure if they are made for racing or for public transportation, but it's nice to know that they're a sports car. I wonder if they are made for public transportation?
- [ ] I'm not sure, but it is a very well known car. I love the Ford Mustang. It is very popular. Do you know anything else about it?
- [ ] Wow, that is crazy! I wonder if they had a cool name for the car? I've never heard of that one. I wonder if Ford had a name for it?

Figure 21: The 20th task of the questionnaire of Experiment 2.

Task 21 out of 40

Topic: Acrophobia
CEFR: b2

Person A: I guess it's all about finding a trustworthy source to experience it, but I don't know, I don't think I could do it. The idea of free-falling is not thrilling to me.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] It would be if it were not for the fact that it has been characterized by an acute sense of self-loathing.
- [ ] I'm not sure, but I know that the feeling of being trapped is pretty scary. I'm afraid of going to the bathroom and not being able to walk!
- [ ] It is a common reaction to certain situations, such as a falling tree.
- [ ] I'm not sure what it is, but I know that the feeling is a sort of "Nexus Wave" that can be overwhelming, like being unable to concentrate, or even having to break up with someone.
- [ ] I can't imagine that, I think it's more like having a bad feeling about your body and feeling the need to be more careful.
- [ ] I think it's just like the people who experience it in the morning. I don't even know if that's true.
- [ ] It's a weird feeling. I don't know how to handle it but I have a friend who is afraid to fall off the cliff.
- [ ] I have to agree with you. I find it interesting that people can suffer from feelings of panic and high levels of anxiety, but not actually experience any of those things.

Figure 22: The 21st task of the questionnaire of Experiment 2.

Task 22 out of 40

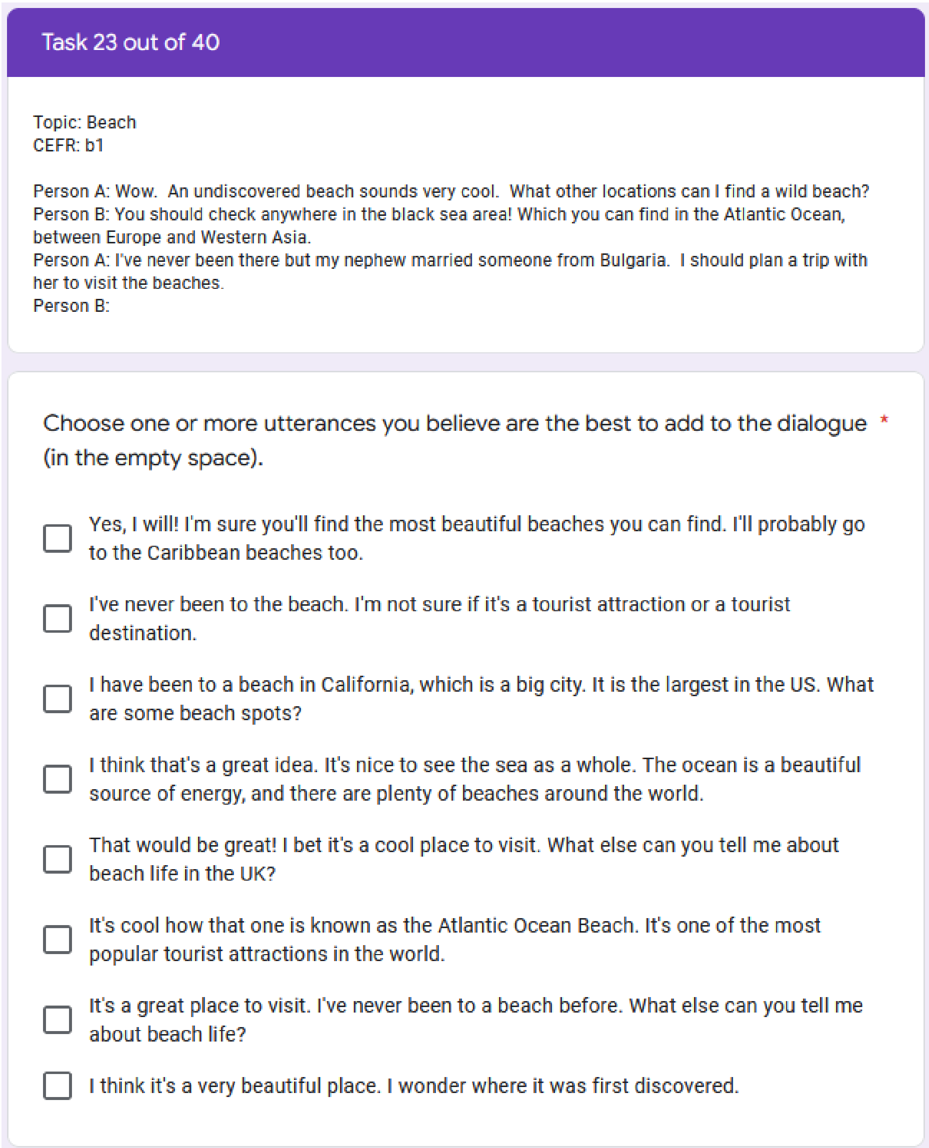
Topic: Acrophobia
CEFR: b2

Person A:
Person B: I guess it's all about finding a trustworthy source to experience it, but I don't know, I don't think I could do it. The idea of free-falling is not thrilling to me.

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] That sounds like fun! I just can't do it! I can't imagine living in a place like that. I don't even know how to learn to read! That's insane! I wonder what would be so hard to do! I'm
- [ ] I'm not sure, what other symptoms do you have? I guess it's a coping mechanism. I am scared of heights.
- [ ] I've heard of that before. I've never been around that before. What's it like to fall? I think it's scary to fall?
- [ ] What would it be like to fall in love with someone? I think I'd be scared to fall in love with someone I don't know.
- [ ] That sounds very interesting, I never knew that! Can you tell me more about it? I wonder if it's just a form of anxiety. It seems like it can be so intense.
- [ ] I think it would be fun to do. I guess I would try to stay calm and don't let myself get too close to people. But at least I'm not going to be scared of people.
- [ ] The most popular thing I can think of is the notion that they can't get into a fight because they're afraid of heights. I don't know the reasons behind that.
- [ ] I'm not sure. Maybe it's a combination of the two. I'm not sure what causes it. I know that people who have acrophobia are more likely to get high and fall in love with someone who is in a relationship.

Figure 23: The 22nd task of the questionnaire of Experiment 2.


Task 23 out of 40

Topic: Beach
CEFR: b1

Person A: Wow. An undiscovered beach sounds very cool. What other locations can I find a wild beach?
Person B: You should check anywhere in the black sea area! Which you can find in the Atlantic Ocean, between Europe and Western Asia.
Person A: I've never been there but my nephew married someone from Bulgaria. I should plan a trip with her to visit the beaches.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes, I will! I'm sure you'll find the most beautiful beaches you can find. I'll probably go to the Caribbean beaches too.
- [ ] I've never been to the beach. I'm not sure if it's a tourist attraction or a tourist destination.
- [ ] I have been to a beach in California, which is a big city. It is the largest in the US. What are some beach spots?
- [ ] I think that's a great idea. It's nice to see the sea as a whole. The ocean is a beautiful source of energy, and there are plenty of beaches around the world.
- [ ] That would be great! I bet it's a cool place to visit. What else can you tell me about beach life in the UK?
- [ ] It's cool how that one is known as the Atlantic Ocean Beach. It's one of the most popular tourist attractions in the world.
- [ ] It's a great place to visit. I've never been to a beach before. What else can you tell me about beach life?
- [ ] I think it's a very beautiful place. I wonder where it was first discovered.

Figure 24: The 23rd task of the questionnaire of Experiment 2.

Task 24 out of 40

Topic: Beach
CEFR: b1

Person A:
Person B: Wow. An undiscovered beach sounds very cool. What other locations can I find a wild beach?
Person A: You should check anywhere in the black sea area! Which you can find in the Atlantic Ocean, between Europe and Western Asia.
Person B: I've never been there but my nephew married someone from Bulgaria. I should plan a trip with her to visit the beaches.

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- ☐ I haven't but I've been to a few places in Florida. It's pretty cool. I've never been there but it is the most beautiful place in the world.
- ☐ Yes, I really like the beaches in the Atlantic Ocean. It is so far from the sea. It's so close to the Pacific Ocean. It's so beautiful! It's a lot of fun. Have you ever been
- ☐ I have never been there. I was just looking for a beach. It's so nice to have a natural environment to explore.
- ☐ I'm not sure, but it is the most popular beach in the US and the second most popular in Europe. Do you know where it is popular?
- ☐ I love beaches! I also love to go to the beach on a weekend. Have you ever been to the beach? I know it's called the Golden Gate, but I've never been there before. Would you mind?
- ☐ I like to go to a beach for the weekend. It's a great place to relax. I like to go to a beach for a long weekend.
- ☐ I've never been to a beach but I've seen a lot of rock formations. I would love to visit the beach.
- ☐ Yes, the beach is so beautiful and interesting. It is a great place to visit! I've never been to a beach before.

Figure 25: The 24th task of the questionnaire of Experiment 2.

Task 25 out of 40

Topic: Immigration to the United States
CEFR: b2

Person A: That's true, and considering the US had nearly open borders until 1924, erecting further barriers seems hypocritical! When their parents benefited from open borders.
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes, and while there is some debate over immigration, a number of countries still do allow it.
- [ ] It sure is, and I am sure that many immigrants were discouraged from migrating to the US.
- [ ] I am glad you're making this point. Immigration to the United States has been a contentious topic for a long time, and it's been debated by the courts for a long time.
- [ ] I agree, but they were still able to get jobs, and move around in America, which is why I'm not surprised they came.
- [ ] Yes, and the US has a wide variety of immigrants and asylum seekers. I think the US is the most populous country in the world!
- [ ] Yeah, they certainly did! I'm sure that immigrants have many different economic and social barriers that keep them from getting a fair chance at a decent life.
- [ ] They did, and they did so without a problem. But in the United States they are being made illegal as immigrants!
- [ ] I agree! Immigrants were brought to the United States as children and were allowed to study in the US as adults.

Figure 26: The 25th task of the questionnaire of Experiment 2.

Task 26 out of 40

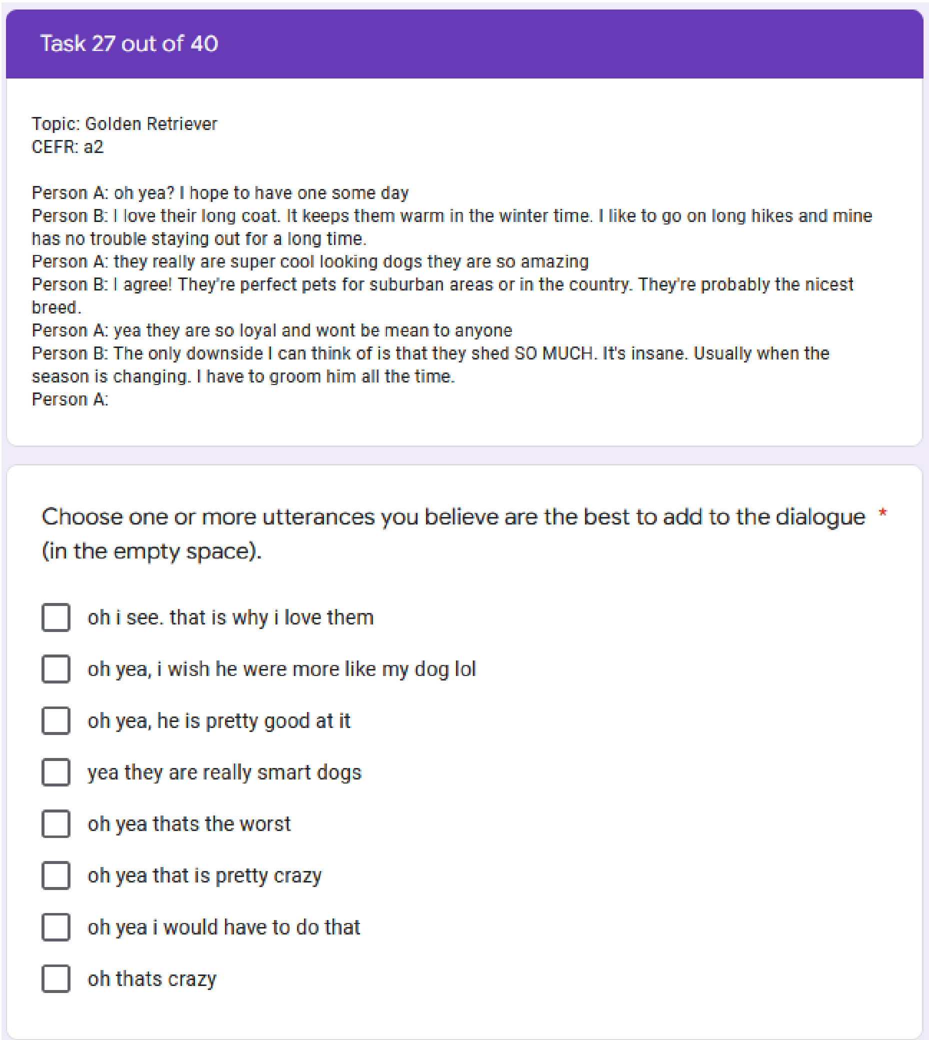
Topic: Immigration to the United States
CEFR: b2

Person A:
Person B: That's true, and considering the US had nearly open borders until 1924, erecting further barriers seems hypocritical! When their parents benefited from open borders.

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes, they are, especially in the South. I have heard that it is the most popular source of immigration to the US.
- [ ] That's true, I have heard about that. They are in the process of becoming a recognized religion. But I'm not sure if they can handle it!
- [ ] I guess it must be hard for them to accept that as a first step in their lives. I'm not sure why they would want to live there.
- [ ] So they didn't really start using the word "immigration" in their names? That's when the US began their policy of "exclusionary" immigration?
- [ ] The US was the first to establish immigration restrictions. It is a huge market.
- [ ] It is. There is no question that immigrants are welcome in the US. It is a huge economy with more than 7 million people in the US. It is also a thriving and vibrant democracy.
- [ ] I'm not sure but we have an estimated population of over 10 million. That's the number of people who have immigrated to the US since 1924!
- [ ] Wow! I bet that's true, we would have a lot more freedom in our country. I wonder if there is a history of migration to the US?

Figure 27: The 26th task of the questionnaire of Experiment 2.



Task 27 out of 40

Topic: Golden Retriever
CEFR: a2

Person A: oh yea? I hope to have one some day
Person B: I love their long coat. It keeps them warm in the winter time. I like to go on long hikes and mine has no trouble staying out for a long time.
Person A: they really are super cool looking dogs they are so amazing
Person B: I agree! They're perfect pets for suburban areas or in the country. They're probably the nicest breed.
Person A: yea they are so loyal and wont be mean to anyone
Person B: The only downside I can think of is that they shed SO MUCH. It's insane. Usually when the season is changing. I have to groom him all the time.
Person A:

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] oh i see. that is why i love them
- [ ] oh yea, i wish he were more like my dog lol
- [ ] oh yea, he is pretty good at it
- [ ] yea they are really smart dogs
- [ ] oh yea thats the worst
- [ ] oh yea that is pretty crazy
- [ ] oh yea i would have to do that
- [ ] oh thats crazy

Figure 28: The 27th task of the questionnaire of Experiment 2.

Task 28 out of 40

Topic: Golden Retriever
CEFR: a2

Person A:
Person B: oh yea? I hope to have one some day
Person A: I love their long coat. It keeps them warm in the winter time. I like to go on long hikes and mine has no trouble staying out for a long time.
Person B: they really are super cool looking dogs they are so amazing
Person A: I agree! They're perfect pets for suburban areas or in the country. They're probably the nicest breed.
Person B: yea they are so loyal and wont be mean to anyone
Person A: The only downside I can think of is that they shed SO MUCH. It's insane. Usually when the season is changing. I have to groom him all the time.

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- ☐ They're so cute and smart, so cute they are so cool. They're a cool breed of dog.
- ☐ I have a Golden Retriever. They're so cute. They're adorable.
- ☐ I think so. They're so cute! I have a Golden Retriever. They're also the most loyal dog breed! They are so sweet! I love them.
- ☐ They are so cute and loving. They're so cute and pretty. It's crazy how they keep their fur and feathers neat and neat.
- ☐ I love Golden Retrievers! They're such a friendly breed! They're so smart and can be trained to be a team. Do you have a favorite breed?
- ☐ I like Golden Retrievers. I have a Golden Retriever and I love them. They're so cute. They're so nice to have around.
- ☐ I don't have a favorite dog. My pet is a golden retriever. I love them. They're a great companion.
- ☐ I have two Golden Retriever's. I have a lot of respect for them. They are such a loyal dog!

Figure 29: The 28th task of the questionnaire of Experiment 2.

Task 29 out of 40

Topic: Peanut
CEFR: a2

Person A: So it a legume- which means that it's very solid and kind of oily. It is like a combination of a garbanzo bean and an almond.
Person B: Interesting I think I will like to see them and taste them
Person A: World production of peanuts was 42 million tonnes just 4 years ago, so you should be able to find some!
Person B: OK what could possibly be the production in the next 2 years
Person A:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Well, the peanut is also known as an alkali plant, which means it has a pretty good taste.
- [ ] Peanuts are actually a type of plant called a "peanut" which is a kind of insectoid insect.
- [ ] The production in the next 2 years is going to be huge. It's not just the production that's going to happen. The peanut is also used in a variety of other foods.
- [ ] I think it's possible to find peanuts on the market, but I am not sure.
- [ ] There are a lot of reasons for that. There are over 1.7 billion peanut farmers in the world.
- [ ] Well, it is known as a variety of peanuts that is used in many food products, such as peanut butter and peanut butter bars.
- [ ] I think the production is likely to start in 2017, as the world's supply of peanuts is increasing.
- [ ] Probably peanuts, but I have not eaten them.

Figure 30: The 29th task of the questionnaire of Experiment 2.

Task 30 out of 40

Topic: Peanut
CEFR: a2

Person A:
Person B: So it a legume- which means that it's very solid and kind of oily. It is like a combination of a garbanzo bean and an almond.
Person A: Interesting I think I will like to see them and taste them
Person B: World production of peanuts was 42 million tonnes just 4 years ago, so you should be able to find some!
Person A: OK what could possibly be the production in the next 2 years

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes and I like the taste of peanut butter
- [ ] So what is the flavor of peanut?
- [ ] I love peanuts but do not like to eat them
- [ ] Oh wow i didnt know that, how much does it take to make a peanut? I've never tried it.
- [ ] That is really interesting I would love to see it. What does it taste like?
- [ ] OK I will like them with some peanuts
- [ ] Interesting I think I will like to eat them, I don't have a whole lot of time for them
- [ ] I do not know the exact number of people that make peanuts. I think I could find one in the next 5 years

Figure 31: The 30th task of the questionnaire of Experiment 2.

Task 31 out of 40

Topic: Classical music
CEFR: b1

Person A: Wow, I didn't know it was that old. Who's the most well-known composer?
Person B: Probably Ludwig van Beethoven. He composed music during what was known as the common-practice period. It spanned from 1550 to 1900!
Person A: What's the current period known as?
Person B: You know, I'm not sure! But I do know that the period from 1750 to 1820 was also known as the Classical period, so there's no true defined timeline.
Person A:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Oh, that's pretty interesting. I wonder how many of these musical styles are still around today?
- [ ] Wow, that's interesting. I didn't know that. What is the classical period?
- [ ] That's interesting. What's the most popular classical music?
- [ ] How does it differ from the common practice period?
- [ ] Wow, I didn't know that. How long did the classical period last?
- [ ] What else can you tell me about classical music?
- [ ] Do you know when it began?
- [ ] That's interesting. Is the classical period the period of the Renaissance or the period of the Renaissance?

Figure 32: The 31st task of the questionnaire of Experiment 2.

Task 32 out of 40

Topic: Classical music
CEFR: b1

Person A:
Person B: Wow, I didn't know it was that old. Who's the most well-known composer?
Person A: Probably Ludwig van Beethoven. He composed music during what was known as the common-practice period. It spanned from 1550 to 1900!
Person B: What's the current period known as?
Person A: You know, I'm not sure! But I do know that the period from 1750 to 1820 was also known as the Classical period, so there's no true defined timeline.

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- [ ] It was actually founded in 1892 by John Bellamy. It was a very popular musical genre. It was a period in which the whole musical spectrum was influenced by classical music.
- [ ] I don't know, but I know it was the work of Johann Wolfgang von Beethoven! I know that the work of his mother was in the 16th century, so it's not as if it's older than that.
- [ ] I'm not sure, but the term classical music was first recorded in the 19th century in England. It's still very much an established genre.
- [ ] Well, it was discovered in the 1750s. It's a little old now, but it was one of the first classical music to be considered an experimental genre.
- [ ] I like Classical music! It's the musical style that originated in Europe and was used in the Renaissance. I love classical music!
- [ ] I'm not sure but I do know that there was a Renaissance period in the US, where classical music was first recorded. It's interesting, since it was recorded from the Renaissance.
- [ ] That's cool! I've heard that it's a very early tradition of classical music. It was first recorded in 1650, when Johann Wolfgang von Goethe was a composer.
- [ ] I'm not sure! But the earliest classical music was composed in the 16th century by Ludwig van Beethoven.

Figure 33: The 32nd task of the questionnaire of Experiment 2.

Task 33 out of 40

Topic: Ocean
CEFR: a2

Person A: For example, the Baltic sea is a sea of the Atlantic Ocean.
Person B: I did not know that - what other seas are there?
Person A: There are a bunch. Another common one is the North Sea, also of the Atlantic Ocean. It's between Great Britian, Scandanavia, Germany, The Netherlands, Belgium and France
Person B: There sure are a lot of them.
Person A: Yes, and many of them are in the Atlantic Ocean, which makes up 20 percent of the earth's surface
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes, that's a lot of water! They also have to be very deep to have any life.
- [ ] Yes, the Atlantic Ocean is the most densely populated part of the world.
- [ ] That is a lot of sea. I wonder what the oceans look like?
- [ ] I think that is a lot of oceans. I bet there are a lot of species of sea birds.
- [ ] I wonder what the oceans are made of.
- [ ] That is a lot of land. I can see why it would be interesting to explore this area.
- [ ] I didn't realize that. It's pretty important for our climate.
- [ ] That's a lot of water. The Atlantic Ocean is the seventh-most active ocean.

Figure 34: The 33rd task of the questionnaire of Experiment 2.

Task 34 out of 40

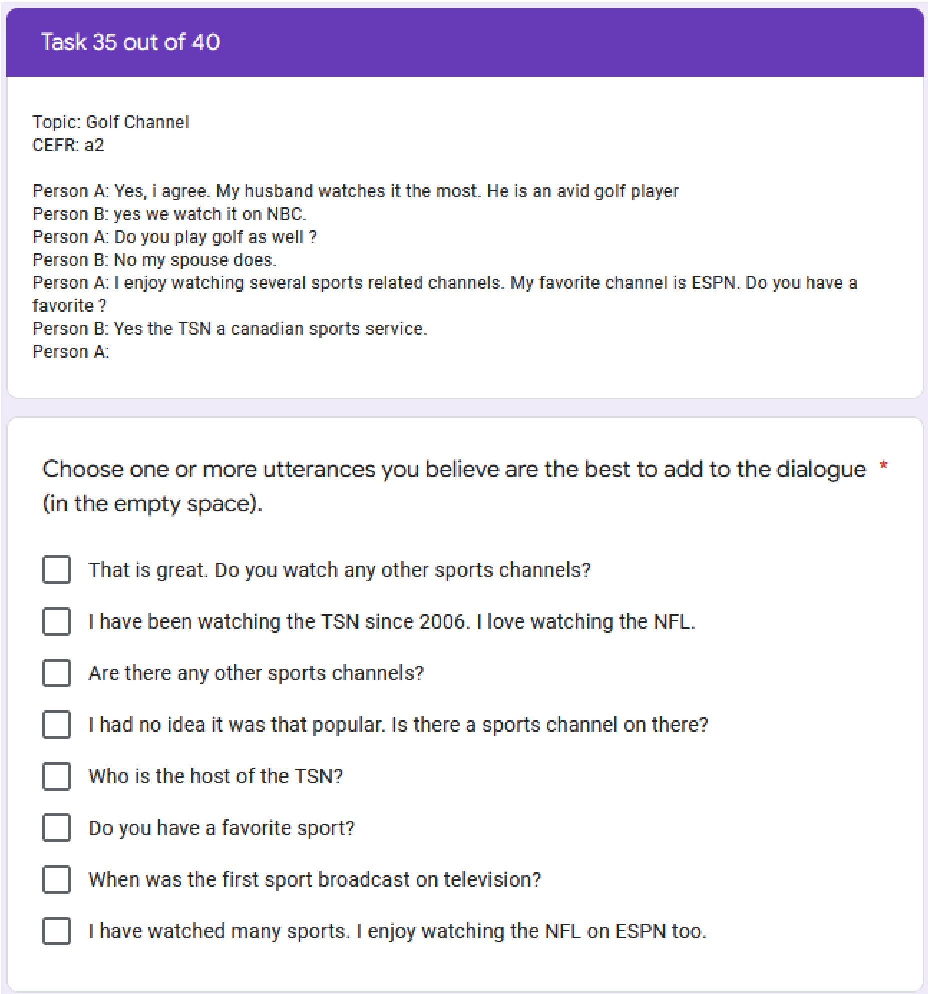
Topic: Ocean
CEFR: a2

Person A:
Person B: For example, the Baltic sea is a sea of the Atlantic Ocean.
Person A: I did not know that - what other seas are there?
Person B: There are a bunch. Another common one is the North Sea, also of the Atlantic Ocean. It's between Great Britian, Scandanavia, Germany, The Netherlands, Belgium and France
Person A: There sure are a lot of them.
Person B: Yes, and many of them are in the Atlantic Ocean, which makes up 20 percent of the earth's surface

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I was thinking of diving into the ocean, but I did not know there was a lot of them!
- [ ] I've never been to the Atlantic Ocean, do you know anything about it?
- [ ] Oh I see. What is a Baltic Sea?
- [ ] I have heard of that. I did not know it was that deep.
- [ ] Oh really? I never knew that! I just know that it's the largest ocean in the world.
- [ ] I do not know much about oceanography, but I do know that the ocean has a variety of forms. Do you know much about it?
- [ ] I have never been to the ocean. I have never been to the ocean. I have never been to the ocean.
- [ ] That is interesting! Where did the ocean originate?

Figure 35: The 34th task of the questionnaire of Experiment 2.



Task 35 out of 40

Topic: Golf Channel
CEFR: a2

Person A: Yes, i agree. My husband watches it the most. He is an avid golf player
Person B: yes we watch it on NBC.
Person A: Do you play golf as well ?
Person B: No my spouse does.
Person A: I enjoy watching several sports related channels. My favorite channel is ESPN. Do you have a favorite ?
Person B: Yes the TSN a canadian sports service.
Person A:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] That is great. Do you watch any other sports channels?
- [ ] I have been watching the TSN since 2006. I love watching the NFL.
- [ ] Are there any other sports channels?
- [ ] I had no idea it was that popular. Is there a sports channel on there?
- [ ] Who is the host of the TSN?
- [ ] Do you have a favorite sport?
- [ ] When was the first sport broadcast on television?
- [ ] I have watched many sports. I enjoy watching the NFL on ESPN too.

Figure 36: The 35th task of the questionnaire of Experiment 2.

Task 36 out of 40

Topic: Golf Channel
CEFR: a2

Person A:
Person B: Yes, i agree. My husband watches it the most. He is an avid golf player
Person A: yes we watch it on NBC.
Person B: Do you play golf as well ?
Person A: No my spouse does.
Person B: I enjoy watching several sports related channels. My favorite channel is ESPN. Do you have a favorite ?
Person A: Yes the TSN a canadian sports service.

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Yes I like to watch golf Channel. Do you watch it on NBC?
- [ ] I love watching golf. Have you ever been to a golf channel?
- [ ] I love watching golf on tv. Have you ever watched Golf Channel?
- [ ] I enjoy watching golf. Do you watch it?
- [ ] I love watching golf Channel. Do you?
- [ ] Hi. I love watching golf. What do you watch?
- [ ] Yes it is! Do you watch it on any channel?
- [ ] Hello, I love watching golf. Do you watch it?

Figure 37: The 36th task of the questionnaire of Experiment 2.

Task 37 out of 40

Topic: Metallica
CEFR: a2

Person A: I've never heard of any of those other bands? Where they as popular as Metallica?
Person B: I think so, especially Anthrax, even though they were the only one of those big four from the east coast.
Person A: What was one of their songs?
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I think the first song was "Black Sabbath" by Led Zeppelin.
- [ ] I'm not sure, but they have a bunch of great songs. They were formed in New York in 1958, when I was a kid.
- [ ] They were the only ones to do anything else, but it was pretty good. They were a very popular band back in the day.
- [ ] They were an American rock band that started in the 1960's and became a huge hit.
- [ ] I'm not sure, but they released a bunch of singles in 2013.
- [ ] It's called "Fearless" and is a song from their debut album, "Loser"
- [ ] It was called "The End of the World" by the band.
- [ ] The one I remember was "Alive" which was pretty good. I think it was pretty good.

Figure 38: The 37th task of the questionnaire of Experiment 2.

Task 38 out of 40

Topic: Metallica
CEFR: a2

Person A:
Person B: I've never heard of any of those other bands? Where they as popular as Metallica?
Person A: I think so, especially Anthrax, even though they were the only one of those big four from the east coast.
Person B: What was one of their songs?

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] Oh yeah, they released their debut album "Loser" back in 2010.
- [ ] They are a very popular band. They have toured the world and have sold over 11 million records!
- [ ] I've heard of them. I like Metallica more. They are very good.
- [ ] I haven't, but I've heard of them. They were pretty good in the 1990s. What was their first album?
- [ ] I'm not sure, but they are known for their music and their dark side. They're great.
- [ ] I have not, but I know they are from the east coast, which is where they originated. I've heard of them.
- [ ] I love Metallica, they're my favorite band.
- [ ] That's true, they have been around for a while, but I'm not sure I've heard of them. Are they still around?

Figure 39: The 38th task of the questionnaire of Experiment 2.

Task 39 out of 40

Topic: The Rolling Stones
CEFR: b1

Person A: I had no idea this invasion had occurred! Are you aware of what type of music they made?
Person B:

Choose one or more utterances you believe are the best to add to the dialogue (in the empty space). *

- [ ] I know that they started in England in the mid-19th century.
- [ ] I am not aware of their music but I know that they were the most successful rock band in history.
- [ ] I am not aware of any of the songs. I'm not sure what the name of one of the members is, but they are considered one of the most influential rock bands in history.
- [ ] I'm not aware of any but they were one of the biggest bands of the 20th century
- [ ] I know that they were a British rock band founded in 1970 and have been around since then.
- [ ] They were a major part of the American rock movement in the late 1950s and early 1960s.
- [ ] I'm not really sure but I know that the founding members of the band included members of The Beatles.
- [ ] They were formed in New York in the 1960s.

Figure 40: The 39th task of the questionnaire of Experiment 2.

Task 40 out of 40

Topic: The Rolling Stones
CEFR: b1

Person A:
Person B: I had no idea this invasion had occurred! Are you aware of what type of music they made?

Choose one or more utterances you believe are the best to add to the dialogue * (in the empty space).

- ☐ No I have never heard of them. I just know that they are a band that began as a British rock band. I have heard of them before. They are pretty good.
- ☐ I know they are from London, England, so that's the number one tour in the UK! They're awesome! I think the Beatles are really famous in the UK.
- ☐ The Rolling Stones were a great rock band! I like their music!
- ☐ I think they have a lot of fun! They were originally formed in 1969. It's very interesting. They've made lots of good music. I don't know much about them. I'll have to check them out.
- ☐ They were quite popular during the 1950s and 1960s! It was such an interesting time!
- ☐ I do! They're a British rock band! I love that band! They're so talented! It's great to hear them play live!
- ☐ I would like to know more about their music! Do you know if the band formed?
- ☐ Well the Rolling Stones are in the UK and have made music in the US. So I guess I'll have to check it out.

Figure 41: The final task of the questionnaire of Experiment 2.

A.2 Experiment 2: Questionnaire Responses

In this appendix, the responses to the questionnaire of Experiment 2 (see Section 5.3) are shown. In each of the figures, the number of participants that picked each of the utterance options is shown.

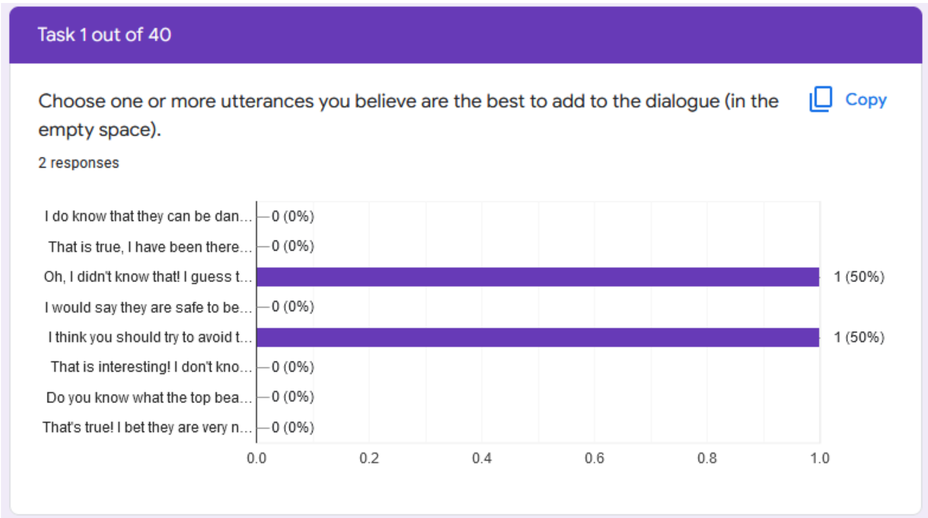


Figure 42: [
Responses to the second task of the questionnaire of Experiment 2]The responses to the first task of the questionnaire of Experiment 2. One of the participants responded that the best utterance to add to the dialogue is the third utterance. The other participant responded that the fifth utterance is the best.

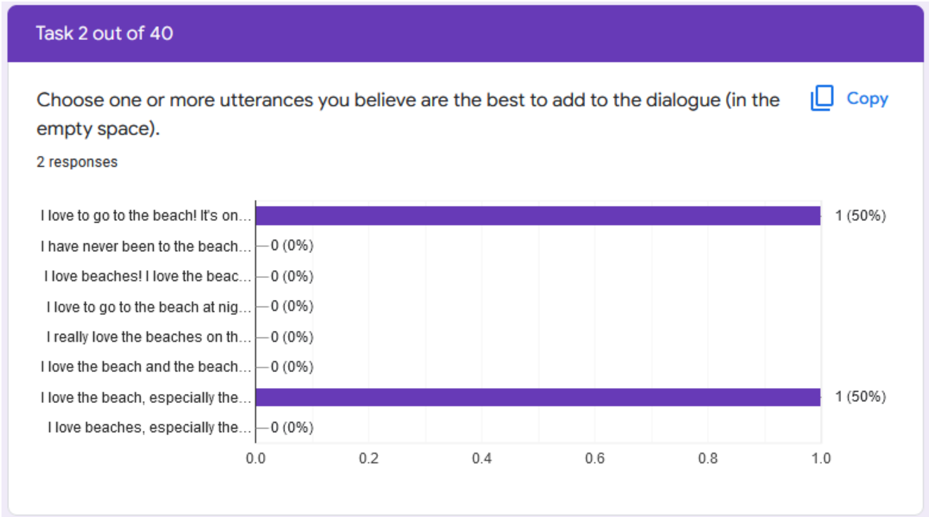


Figure 43: [
Responses to the second task of the questionnaire of Experiment 2]The responses to the second task of the questionnaire of Experiment 2.

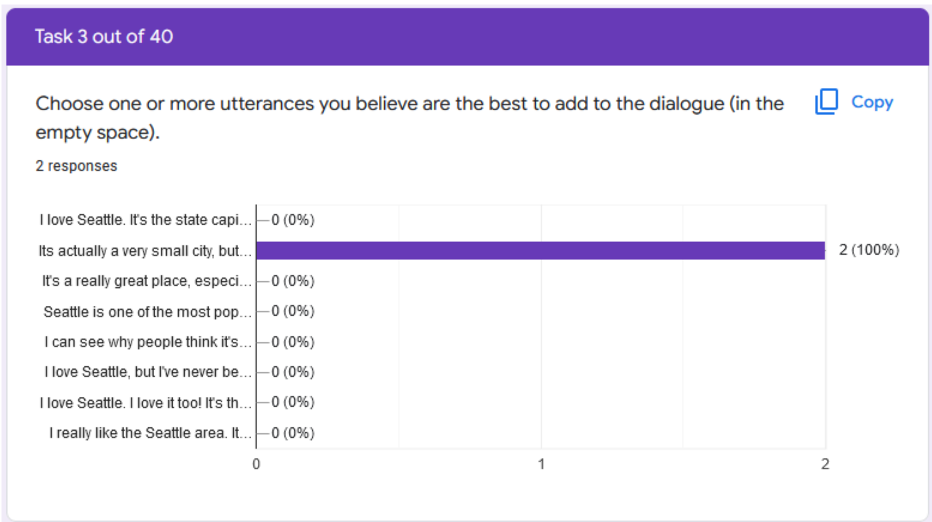


Figure 44: The responses to the third task of the questionnaire of Experiment 2.

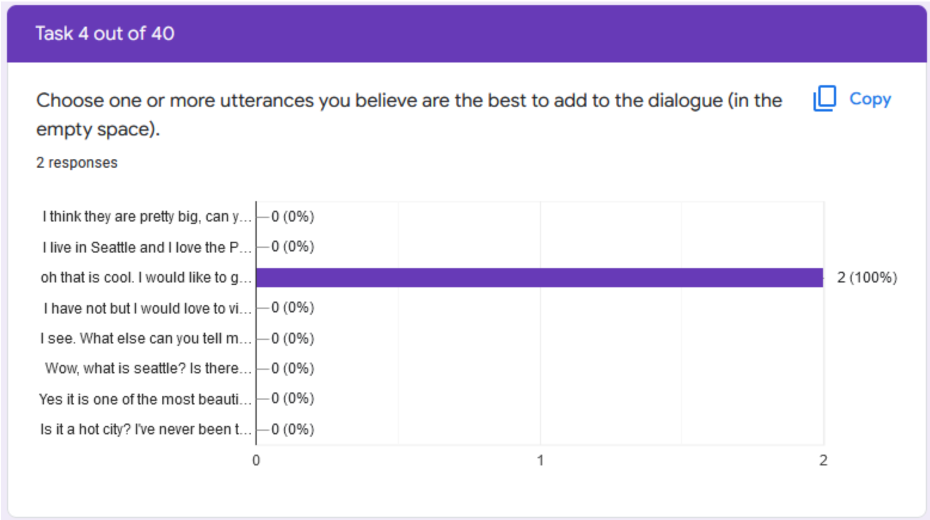


Figure 45: The responses to the fourth task of the questionnaire of Experiment 2.

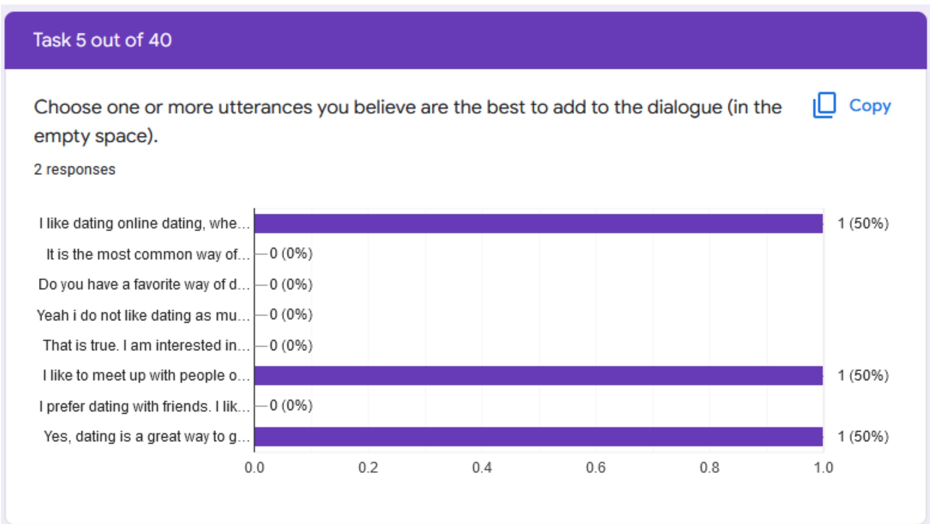


Figure 46: The responses to the fifth task of the questionnaire of Experiment 2.

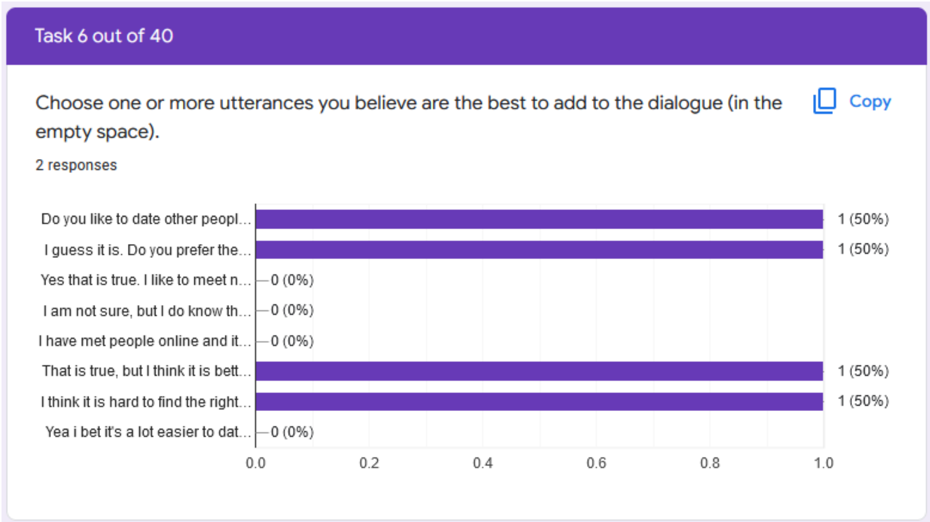


Figure 47: The responses to the sixth task of the questionnaire of Experiment 2.

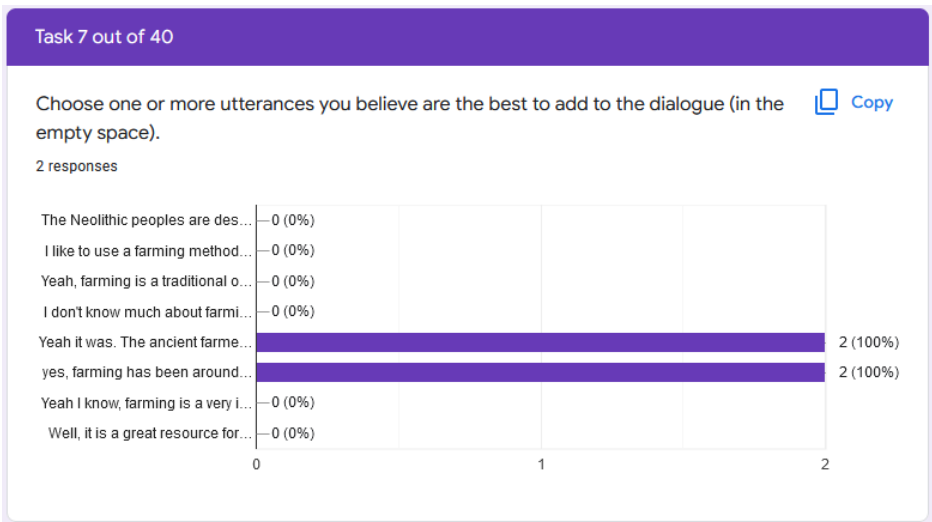


Figure 48: The responses to the seventh task of the questionnaire of Experiment 2.

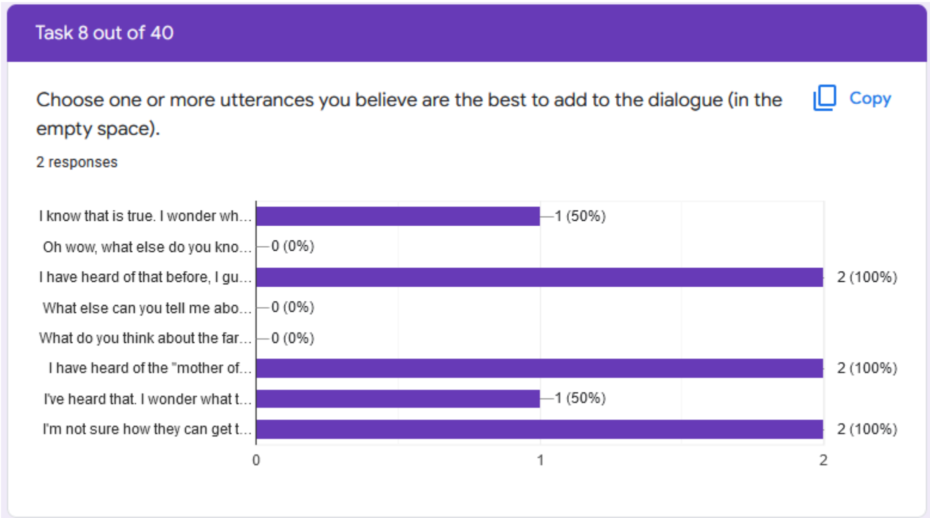


Figure 49: The responses to the eighth task of the questionnaire of Experiment 2.

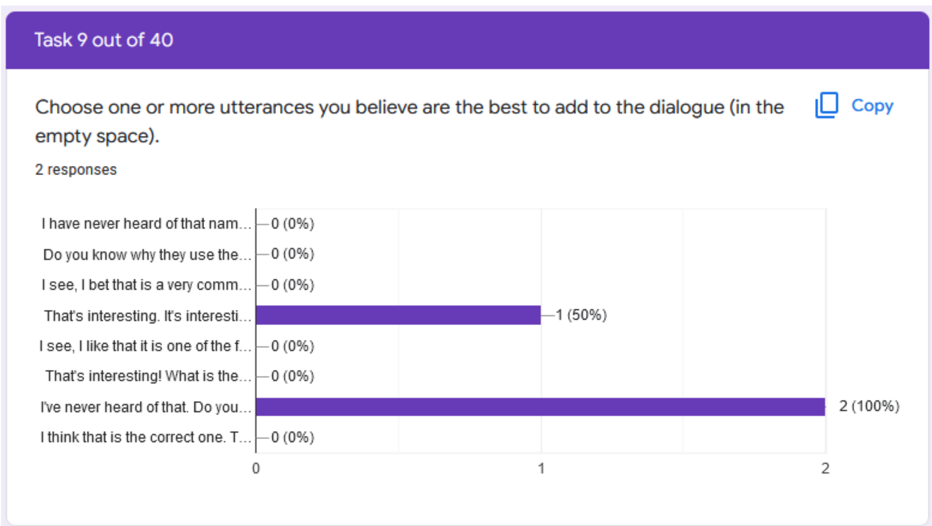


Figure 50: The responses to the ninth task of the questionnaire of Experiment 2.

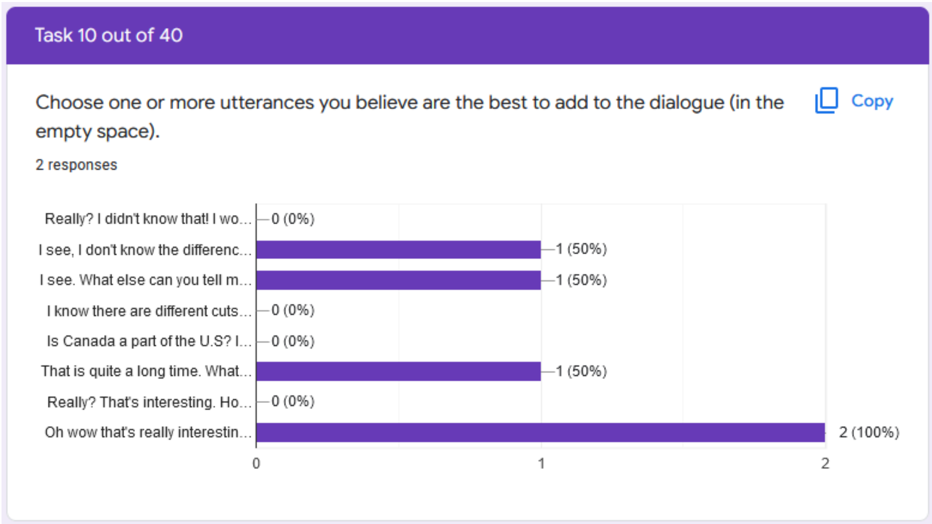


Figure 51: The responses to the 10th task of the questionnaire of Experiment 2.

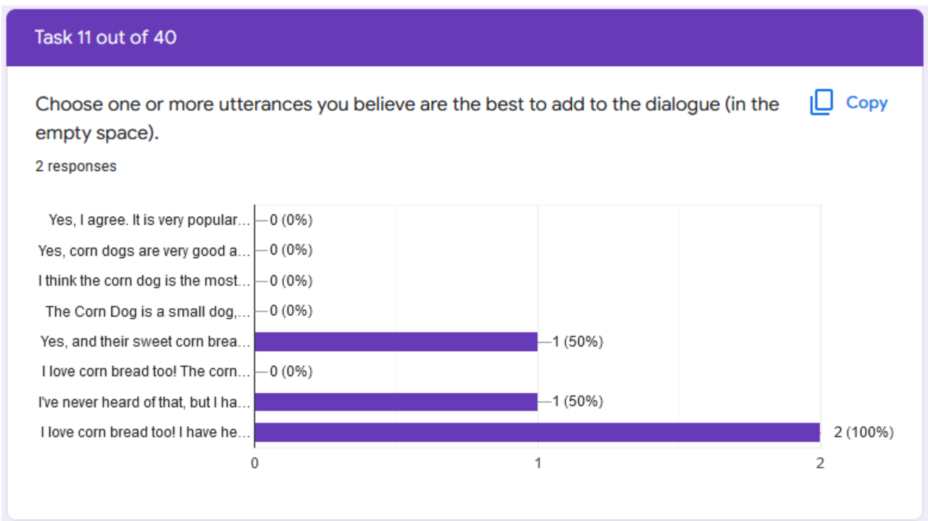


Figure 52: The responses to the 11th task of the questionnaire of Experiment 2.

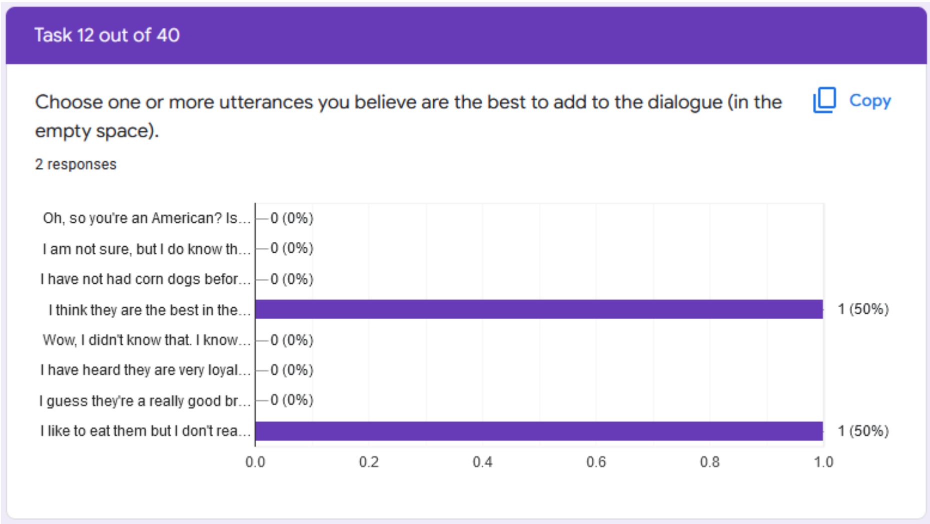


Figure 53: The responses to the 12th task of the questionnaire of Experiment 2.

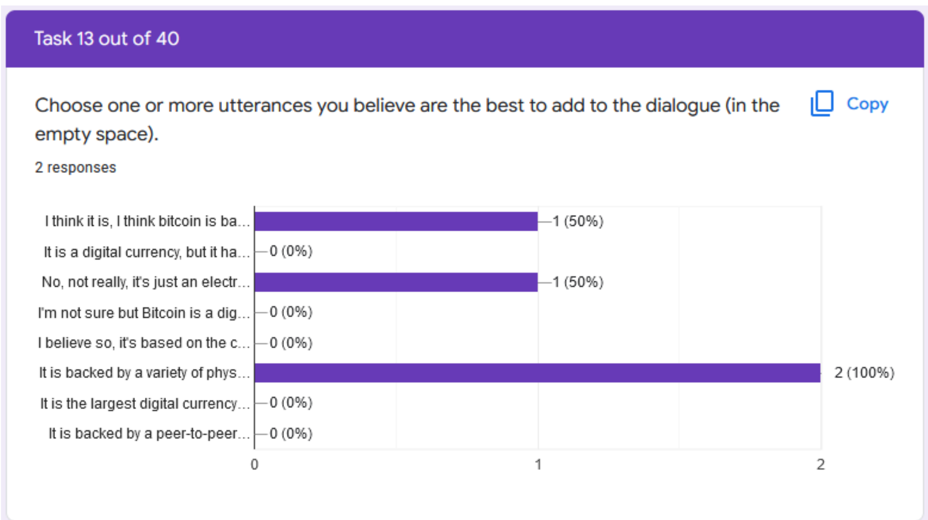


Figure 54: The responses to the 13th task of the questionnaire of Experiment 2.

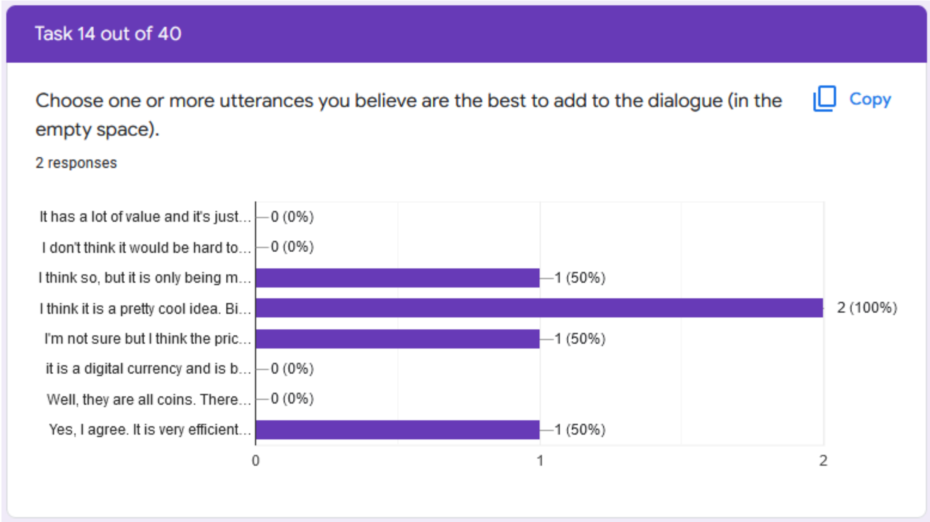


Figure 55: The responses to the 14th task of the questionnaire of Experiment 2.

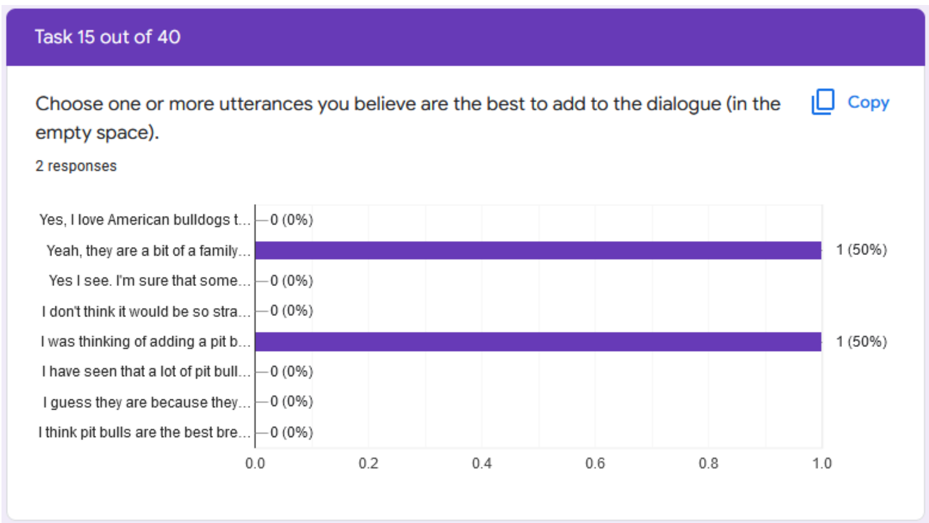


Figure 56: The responses to the 15th task of the questionnaire of Experiment 2.

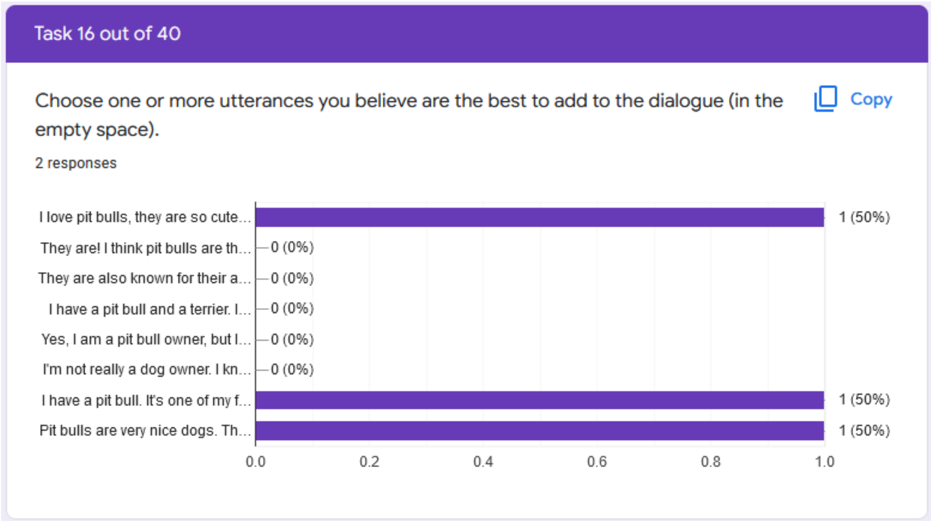


Figure 57: The responses to the 16th task of the questionnaire of Experiment 2.

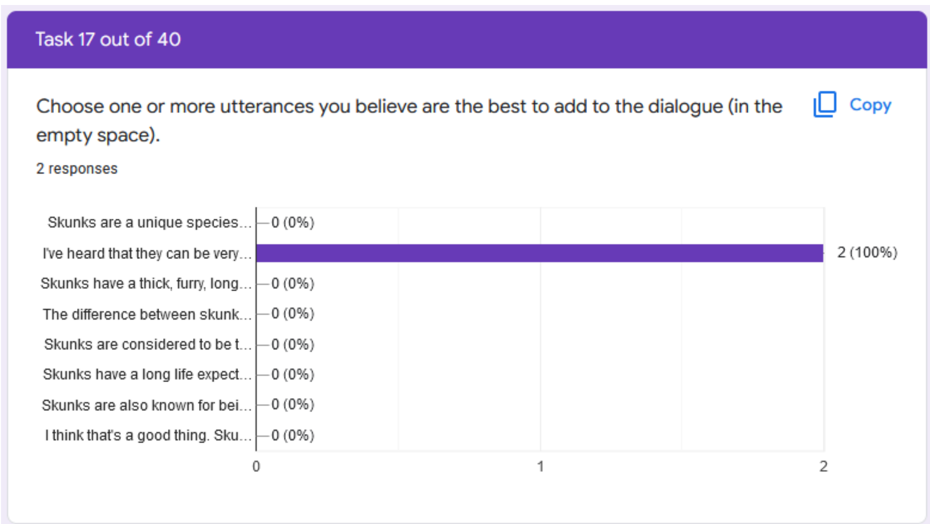


Figure 58: The responses to the 17th task of the questionnaire of Experiment 2.

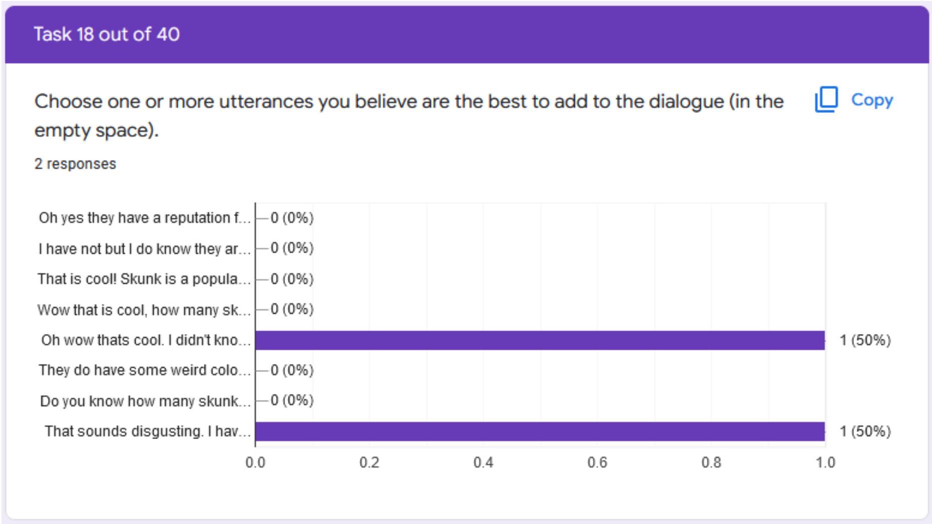


Figure 59: The responses to the 18th task of the questionnaire of Experiment 2.

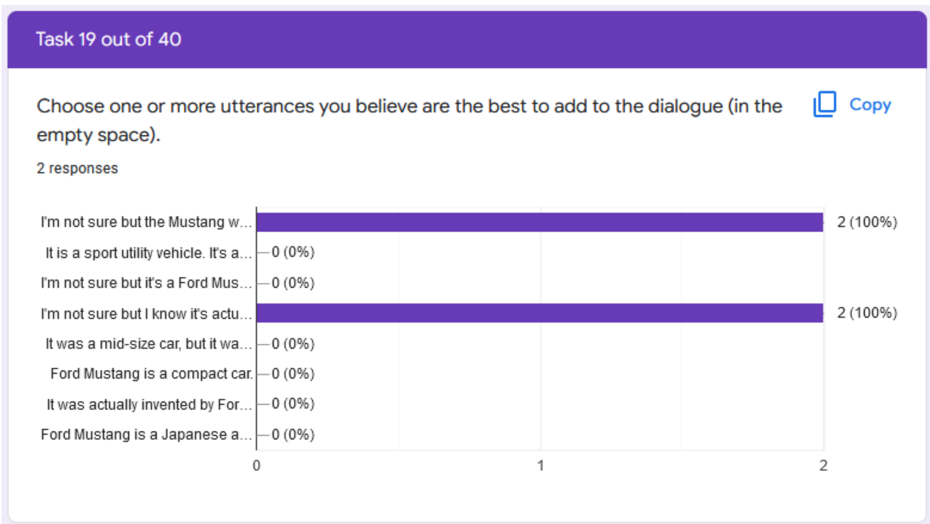


Figure 60: The responses to the 19th task of the questionnaire of Experiment 2.

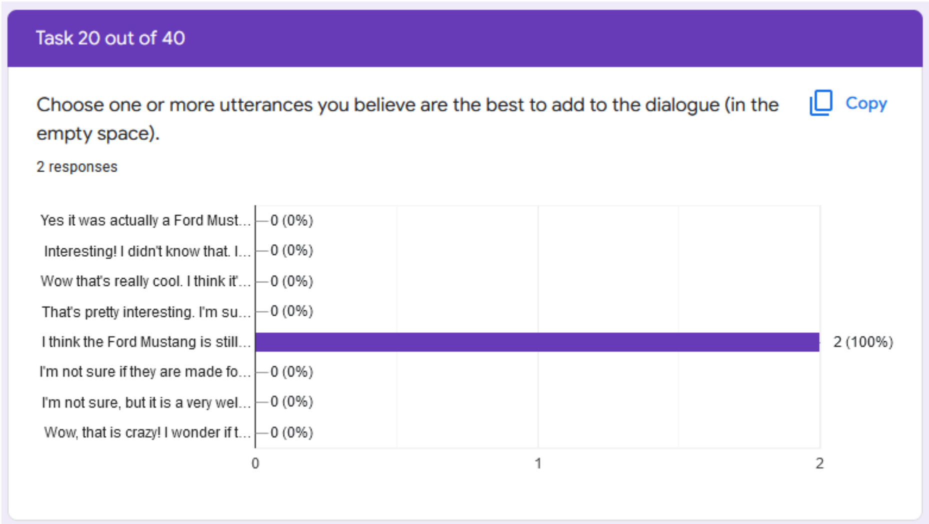


Figure 61: The responses to the 20th task of the questionnaire of Experiment 2.

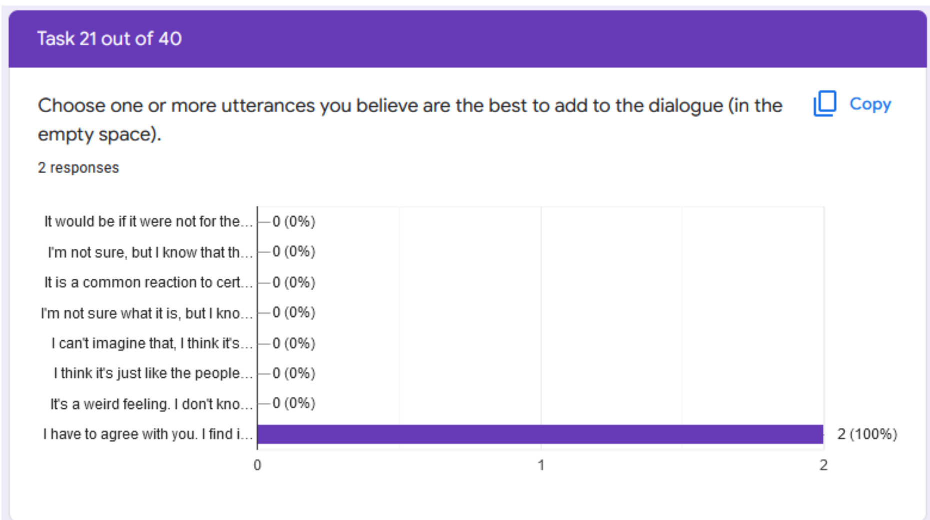


Figure 62: The responses to the 21st task of the questionnaire of Experiment 2.

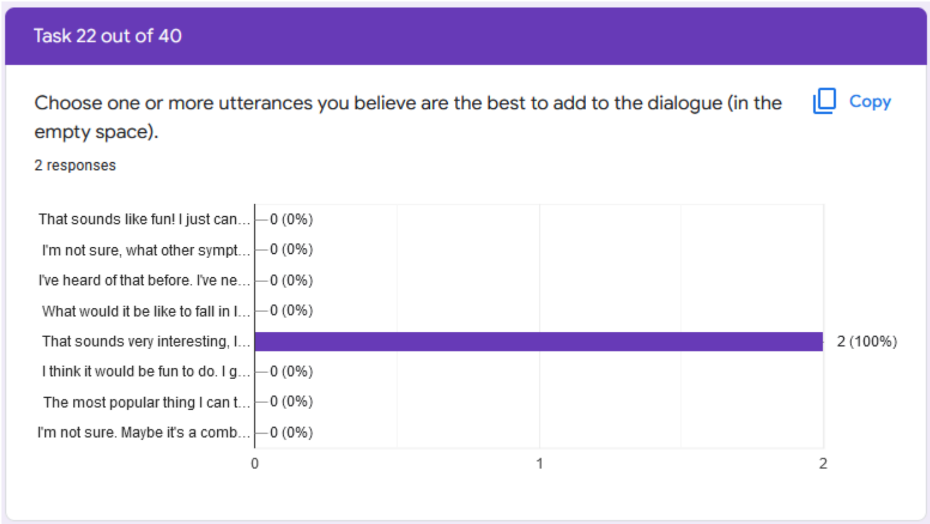


Figure 63: The responses to the 22nd task of the questionnaire of Experiment 2.

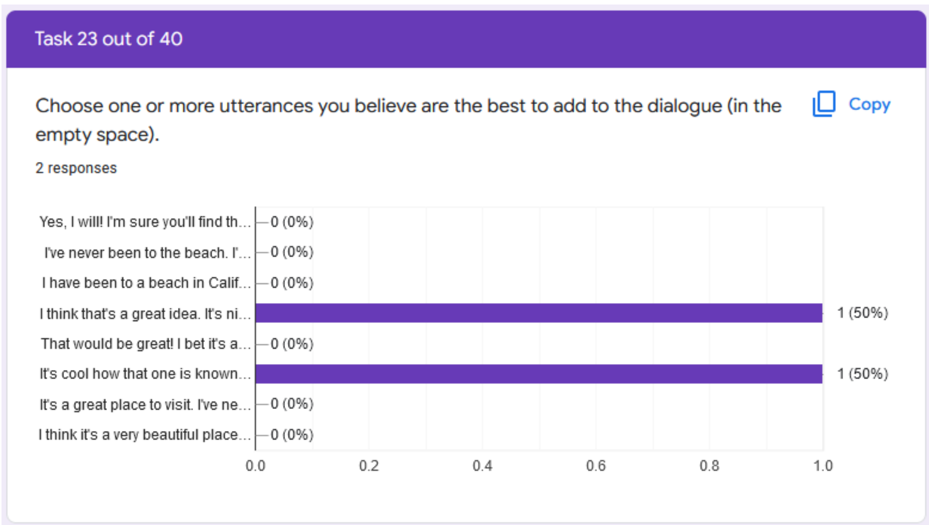


Figure 64: The responses to the 23rd task of the questionnaire of Experiment 2.

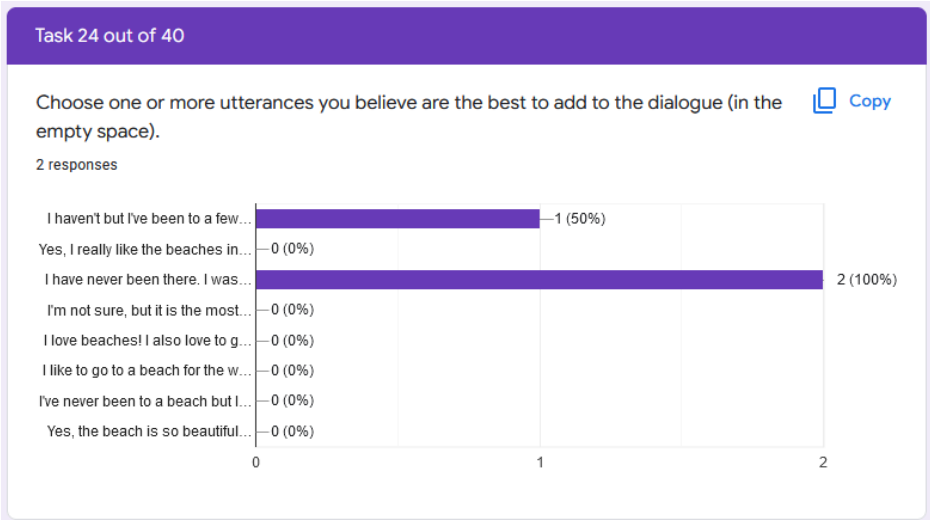


Figure 65: The responses to the 24th task of the questionnaire of Experiment 2.

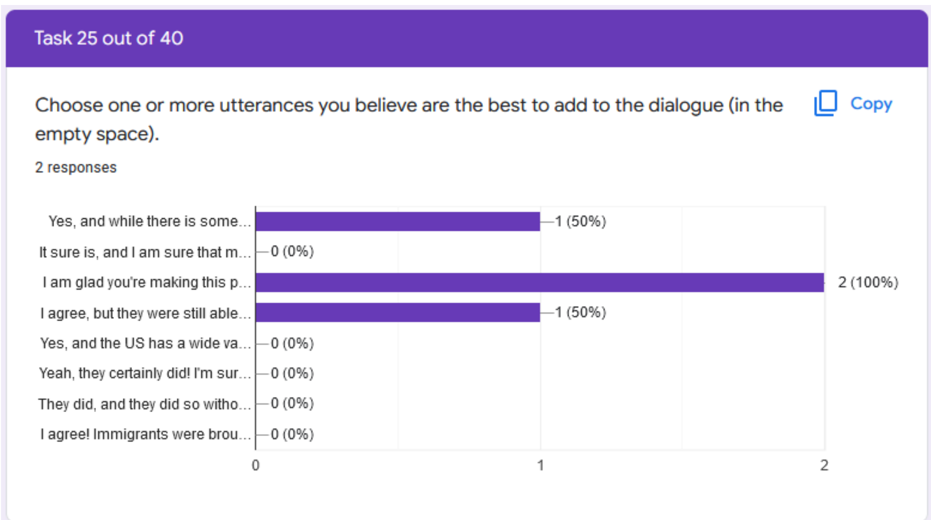


Figure 66: The responses to the 25th task of the questionnaire of Experiment 2.

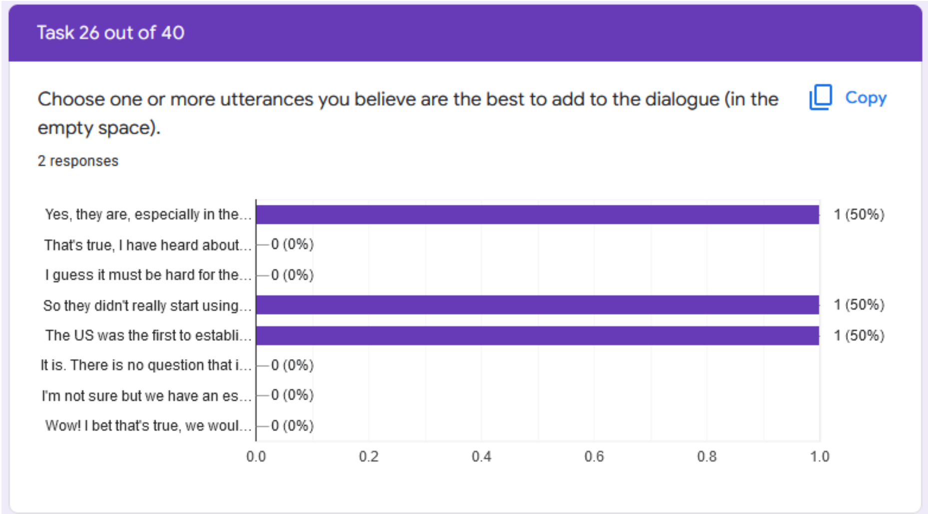


Figure 67: The responses to the 26th task of the questionnaire of Experiment 2.

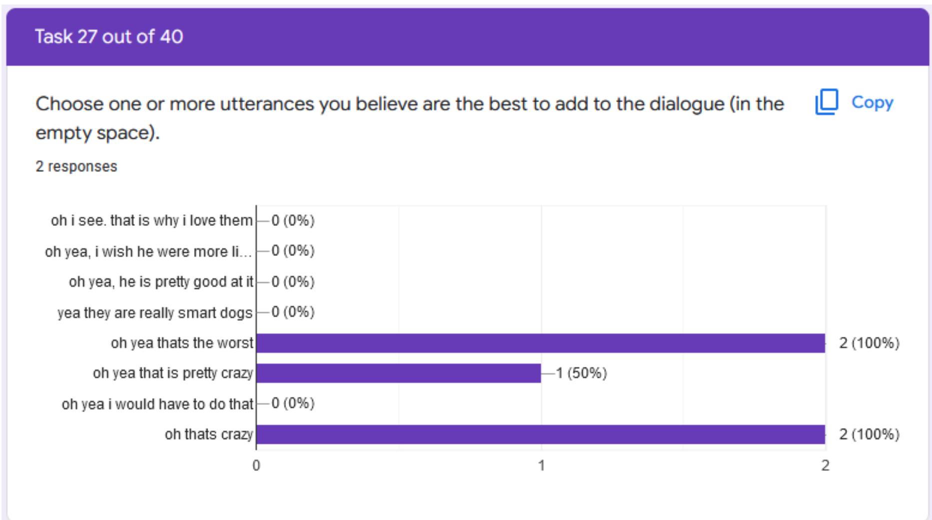


Figure 68: The responses to the 27th task of the questionnaire of Experiment 2.

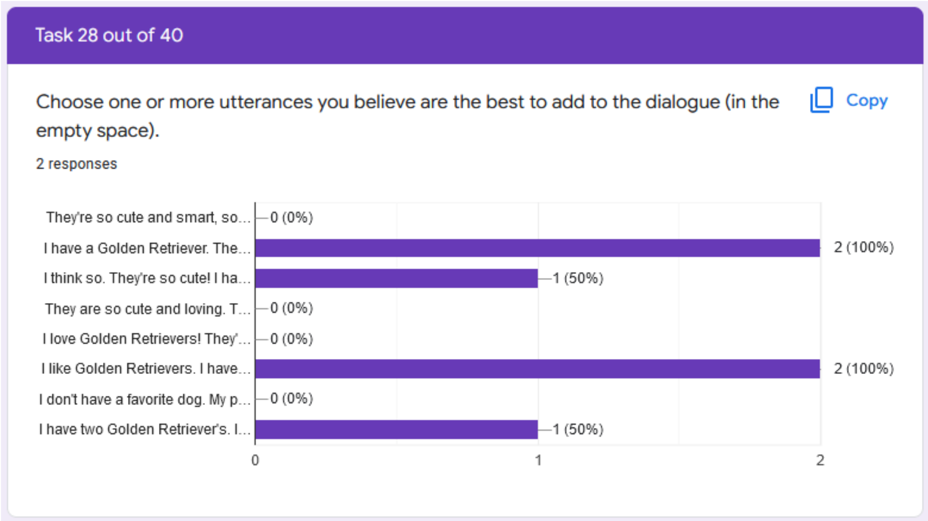


Figure 69: The responses to the 28th task of the questionnaire of Experiment 2.

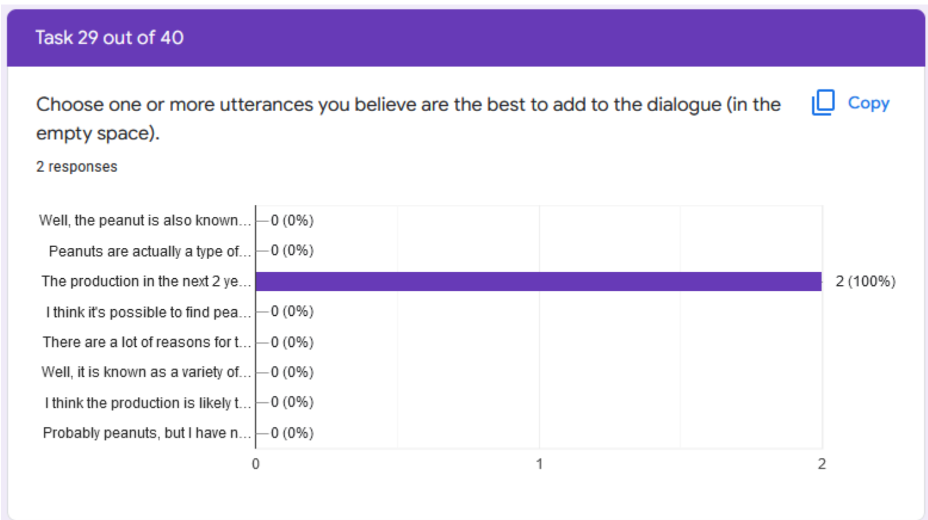


Figure 70: The responses to the 29th task of the questionnaire of Experiment 2.

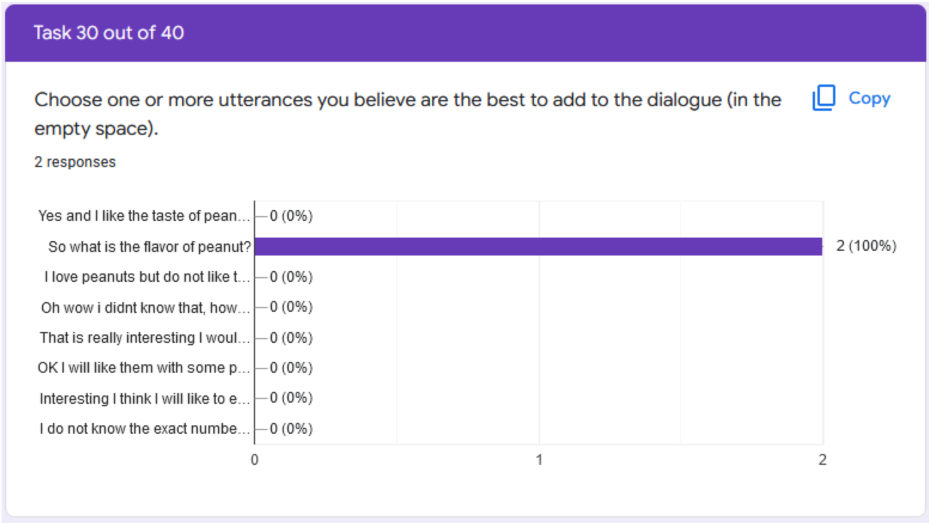


Figure 71: The responses to the 30th task of the questionnaire of Experiment 2.

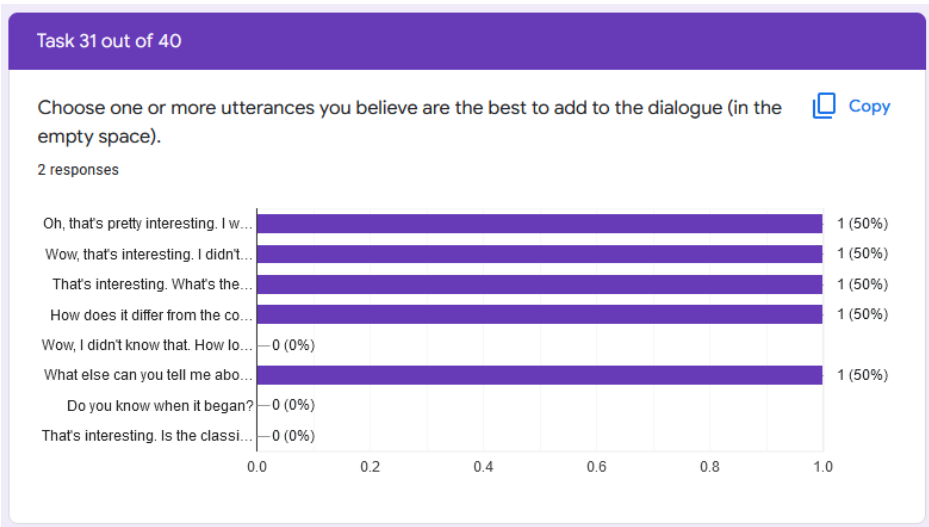


Figure 72: The responses to the 31st task of the questionnaire of Experiment 2.

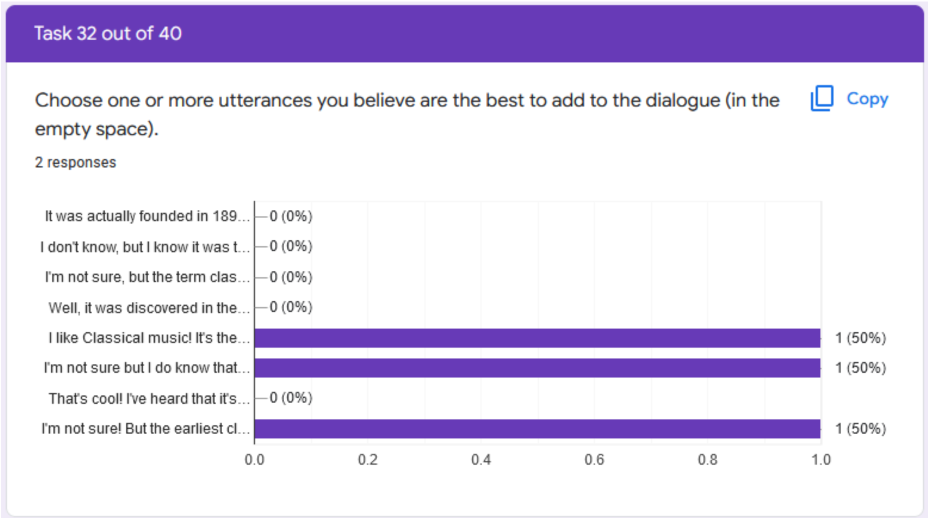


Figure 73: The responses to the 32nd task of the questionnaire of Experiment 2.

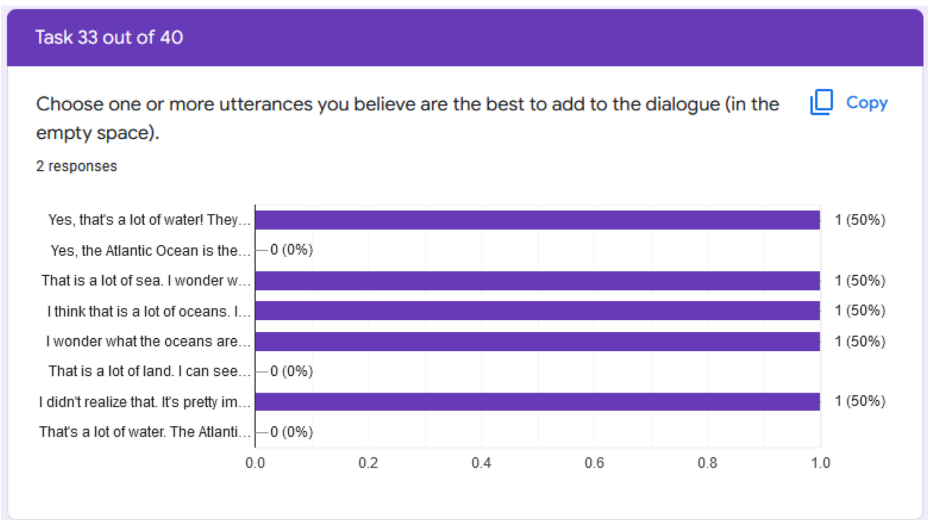


Figure 74: The responses to the 33rd task of the questionnaire of Experiment 2.

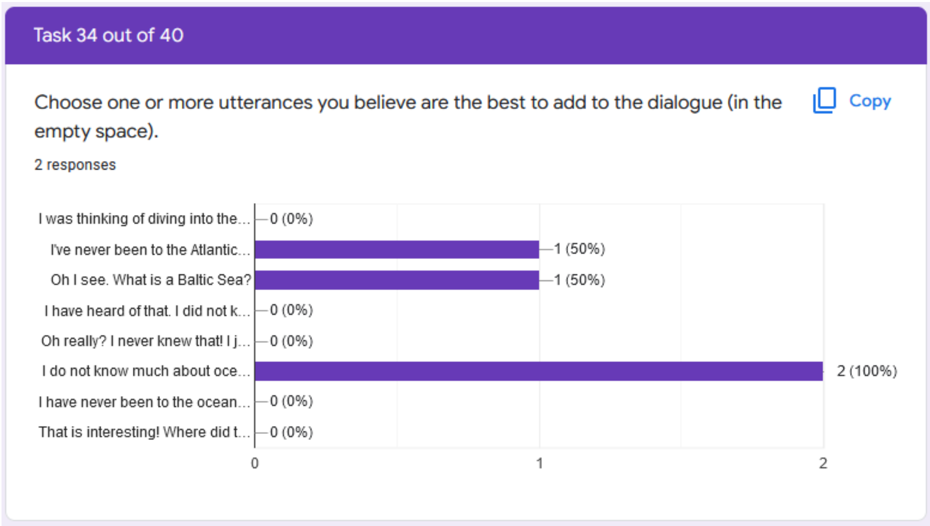


Figure 75: The responses to the 34th task of the questionnaire of Experiment 2.

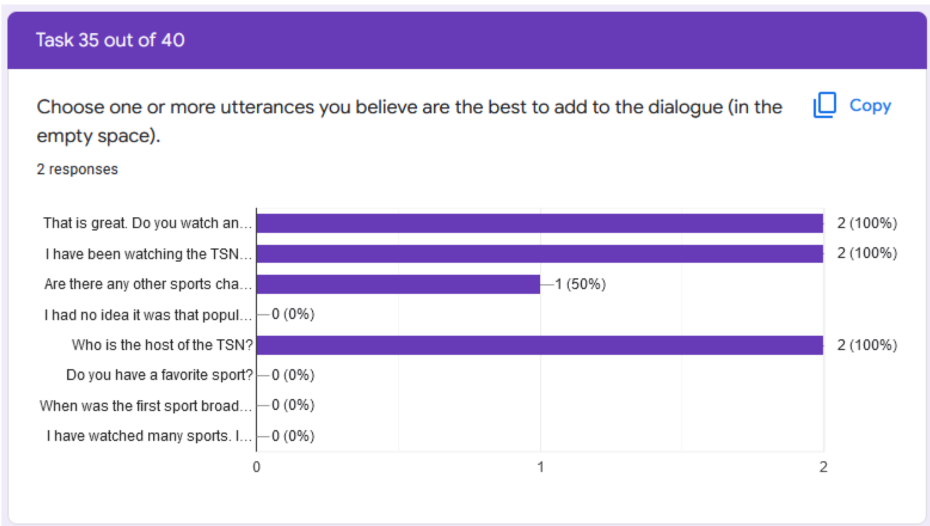


Figure 76: The responses to the 35th task of the questionnaire of Experiment 2.

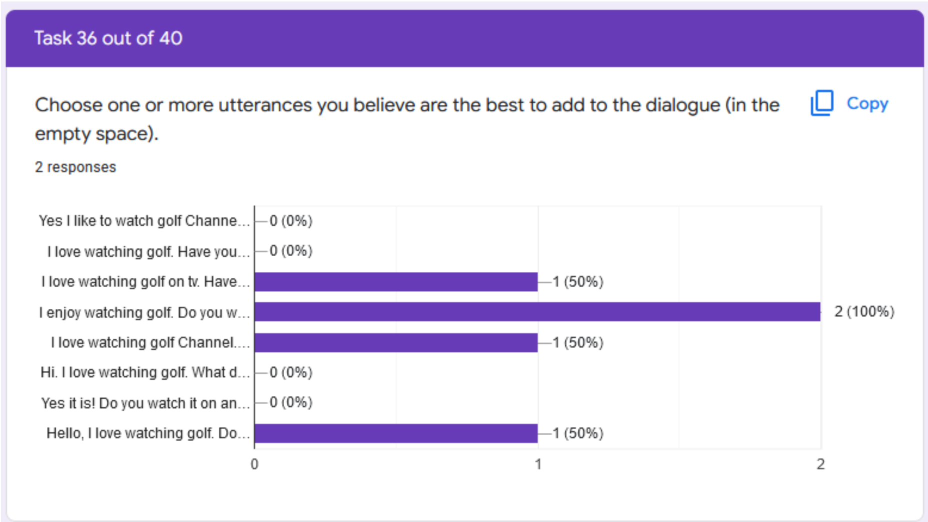


Figure 77: The responses to the 36th task of the questionnaire of Experiment 2.

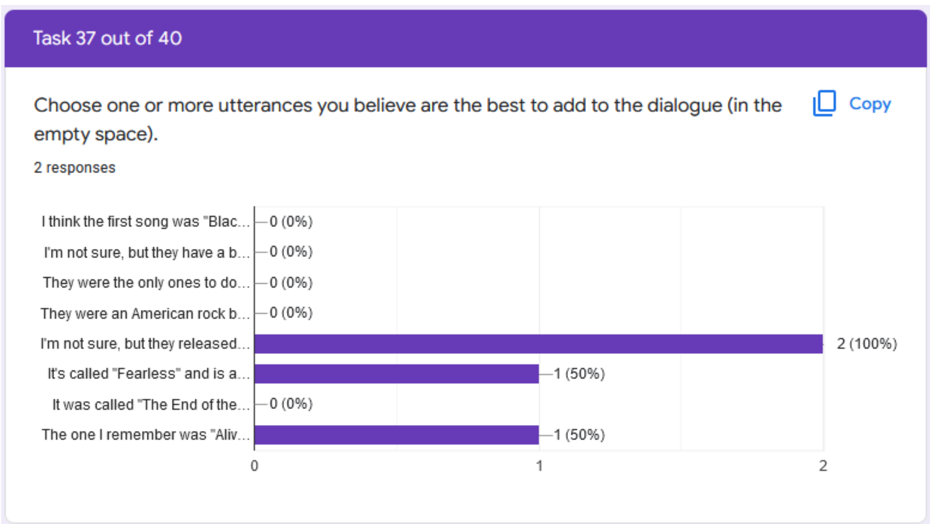


Figure 78: The responses to the 37th task of the questionnaire of Experiment 2.

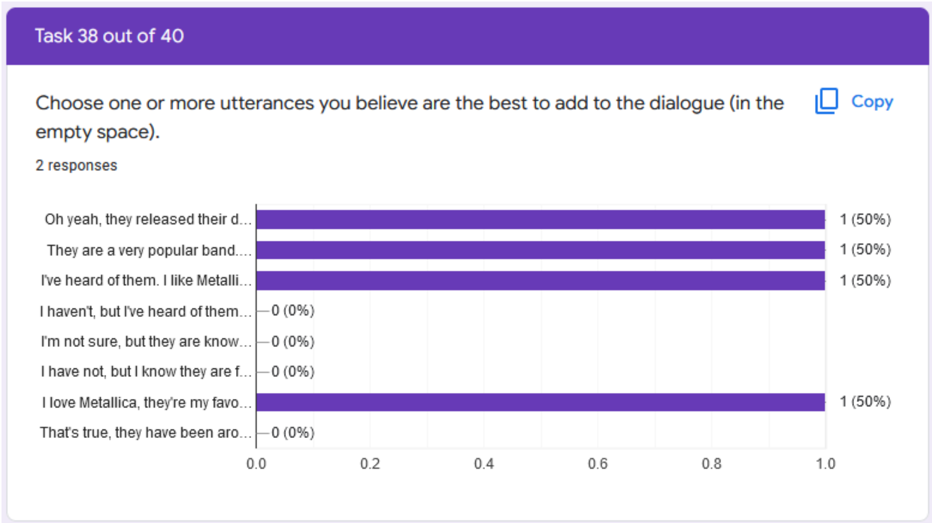


Figure 79: The responses to the 38th task of the questionnaire of Experiment 2.

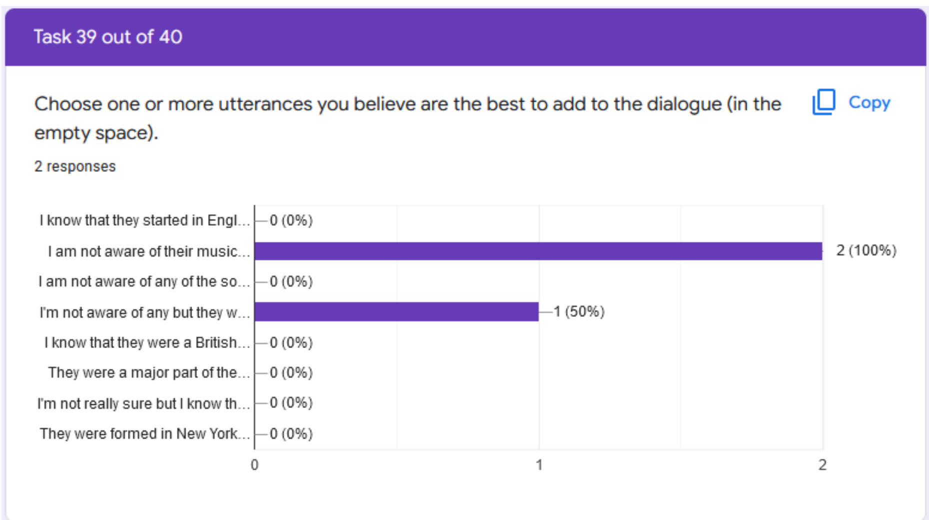


Figure 80: The responses to the 39th task of the questionnaire of Experiment 2.

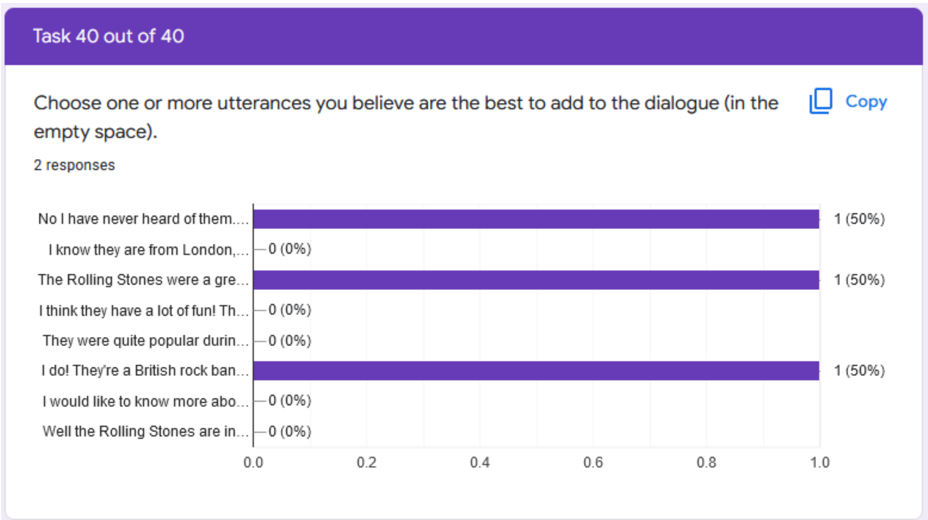


Figure 81: The responses to the final task of the questionnaire of Experiment 2.

A.3 Experiment 3: User Study Instructions

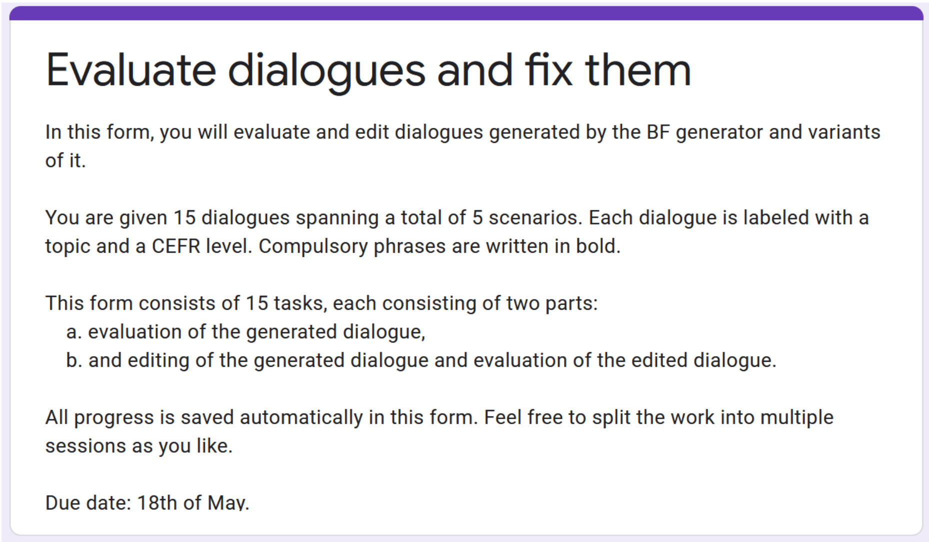
In this appendix, the instructions given to the participants of the questionnaire as part of Experiment 3 (see Section 5.4) are shown, as well as an example of a task they are given. All of the tasks in the questionnaire follow the same structure. Their only difference is the dialogue (and its topic and CEFR level) in question. The dialogues are available in Appendix A.4.5, Appendix A.4.6, and Appendix A.4.7. The instructions on how to write dialogue snippets are shown in Figure 82. The introduction to the questionnaire is shown in Figure 83. A task of the user study is shown in the sections following, with the evaluation of a extended dialogue (see Appendix A.3.2), adjustments of the extended dialogue (see Appendix A.3.3), and evaluation of the extended dialogue (see Appendix A.3.4).

A.3.1 Part 1: Create Dialogue Snippets

Task 1: Create prompts			**Picked**
Create four prompts to the surrounding dialogue generator. A prompt consists of a **Topic**, **CEFR level**, and **Dialogue Snippet**. All dialogue must be in English and between two speakers. Each prompt must consist of 1 to 5 utterances (but keep in mind that each utterance can contain multiple sentences).			
The goal of each dialogue is to teach a user how to speak about the topic on English (given the CEFR level).			
Dialogue Snippet is a snippet of a dialogue that you want to be part of the final dialogue. If multiple utterances are given, the utterances will appear together in the generated dialogue. Please separate utterances with commas.			
The grey rows are examples of prompts.			
Content Creator 1 should fill in the red rows.			
Content Creator 2 should fill in the yellow rows.			
Topic	**CEFR level**	**Dialogue Snippet**	
Informatics	B1	"I found the bug! Finally!"	
Football	A2	"It was so embarrassing today.", "What happened?", "I was the keeper today and got hit straight in my face!"	
Animals	A1	"I have a cat, do you have a cat?", "No, I don't have a cat, I have a dog.", "OK, I don't like dogs.", "Why not?"	x
Food	A2	"Are you hungry?", "Yes, I'm very hungry.", "What do you want to eat?", "I would like a hamburger, please.", "What about you?"	
Spare time activities	B1	"What do you like to do in your spare time?", "I like to play football and read, what about you?", "I like to play the piano and to cook", "What don't you like doing in your spare time?"	
Interests	B2	"I have just finished reading a book about agriculture", "Was it any good?", "Yes, I learned a lot from it, you should read it", "I'm not that interested in agriculture.", "Why not?"	x
Greetings	A1	"Hello, how are you today?", "Hello, I am fine, how are you?", "I am a little tired.", "Why are you tired?"	
Work	A2	"How is work?", "It is great, I love my job!", "What do you work with?", "I work as a teacher. How about you?"	x
Travel/Holiday	B1	"What are your plans for the summer holiday?" "We are going to Greece. I can't wait!", "That´s great, I love Greece! Have you been there before"? "No, it's my first time. Do you have any recommendations?"	x

Figure 82: The instructions given to the participants of the Experiment 3 on how to create input prompts for the surrounding dialogue generator. Each prompt consists of a topic, CEFR level, and dialogue snippet (as seen in the purple row). Two examples (in the grey rows) are given to the content creators for illustrative purposes. The prompts selected to generate extended dialogues for are marked in the right-most column (under "Picked").

A.3.2 Part 2: Evaluate the Extended Dialogues



Evaluate dialogues and fix them

In this form, you will evaluate and edit dialogues generated by the BF generator and variants of it.

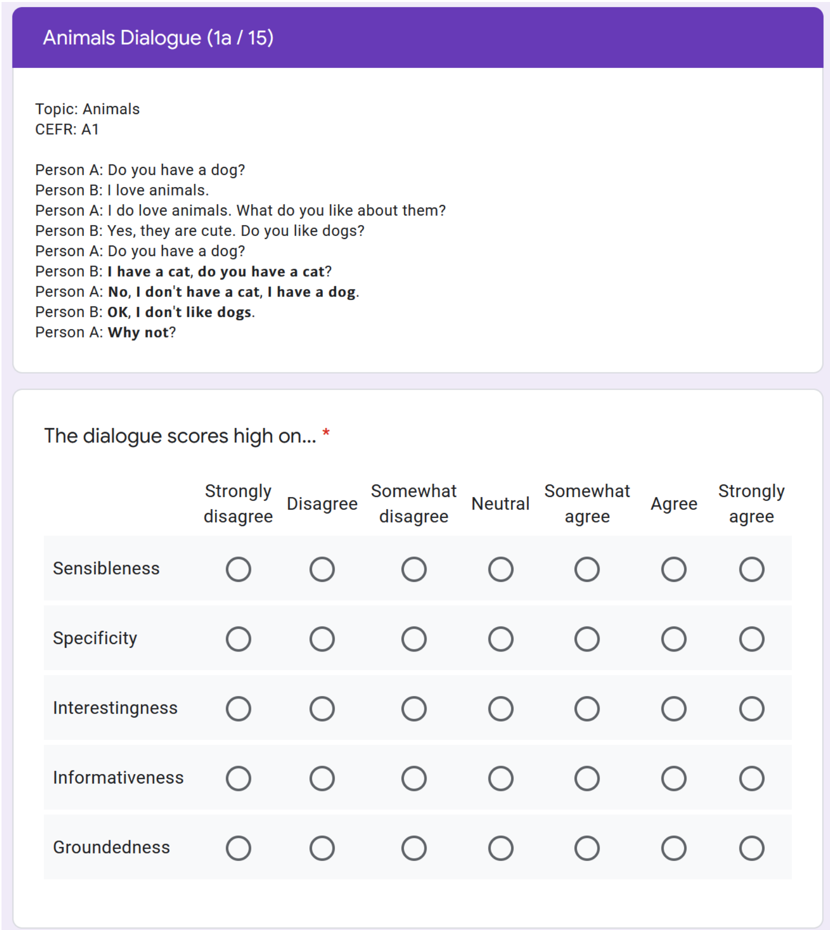
You are given 15 dialogues spanning a total of 5 scenarios. Each dialogue is labeled with a topic and a CEFR level. Compulsory phrases are written in bold.

This form consists of 15 tasks, each consisting of two parts:
a. evaluation of the generated dialogue,
b. and editing of the generated dialogue and evaluation of the edited dialogue.

All progress is saved automatically in this form. Feel free to split the work into multiple sessions as you like.

Due date: 18th of May.

Figure 83: The first instructions given to the participants of the questionnaire of Experiment 3. It gives an overview of the questionnaire.



Animals Dialogue (1a / 15)

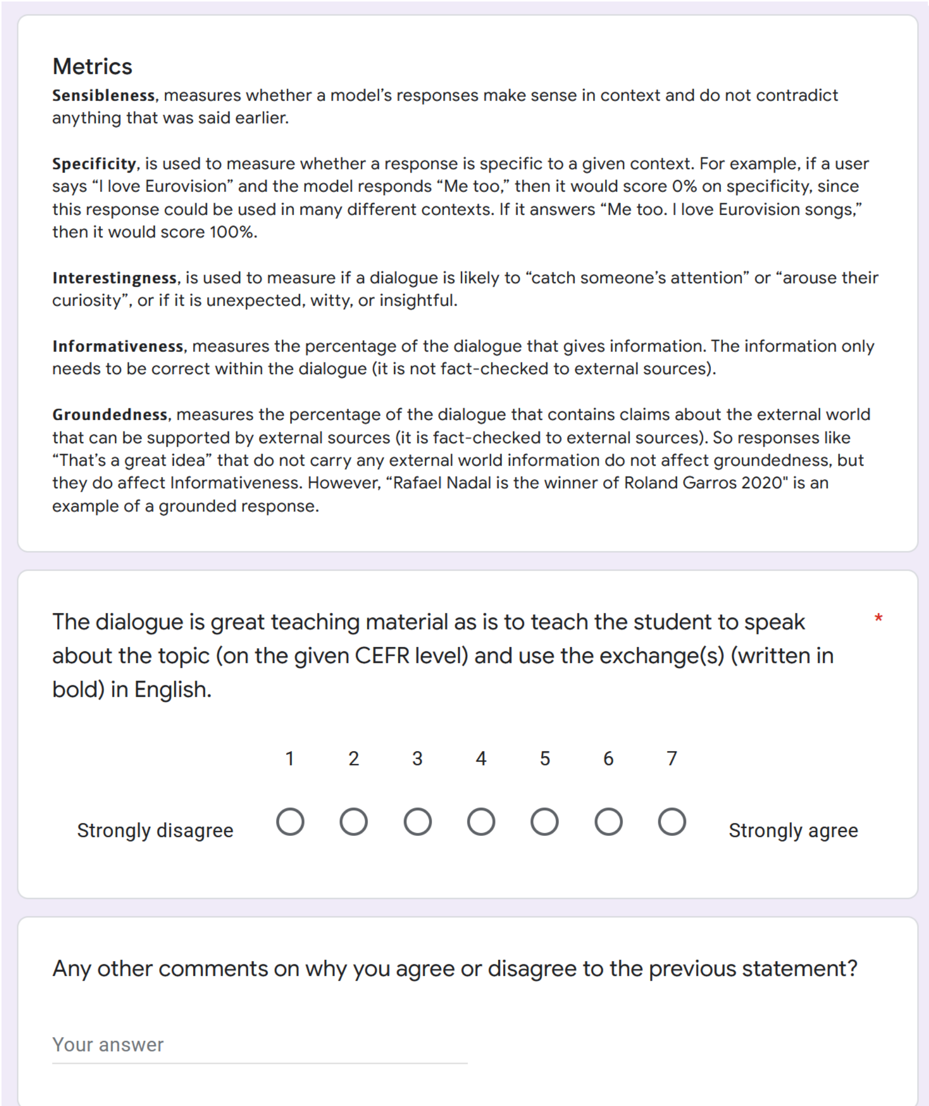
Topic: Animals
CEFR: A1

Person A: Do you have a dog?
Person B: I love animals.
Person A: I do love animals. What do you like about them?
Person B: Yes, they are cute. Do you like dogs?
Person A: Do you have a dog?
Person B: **I have a cat, do you have a cat**?
Person A: **No, I don't have a cat, I have a dog**.
Person B: **OK, I don't like dogs**.
Person A: **Why not**?

The dialogue scores high on... *

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
Sensibleness	○	○	○	○	○	○	○
Specificity	○	○	○	○	○	○	○
Interestingness	○	○	○	○	○	○	○
Informativeness	○	○	○	○	○	○	○
Groundedness	○	○	○	○	○	○	○

Figure 84: The first task of the questionnaire of Experiment 3. A snippet of a dialogue (on the top), including its topic (i.e., Animals) and CEFR level (i.e., A1) is shown. The task is to evaluate the extended dialogue (on the bottom) on the five metrics (i.e., sensibleness, specificity, interestingness, informativeness, and groundedness) on a 7 point Likert scale. By scrolling down on the questionnaire, the participant sees Figure 85.



Metrics

Sensibleness, measures whether a model's responses make sense in context and do not contradict anything that was said earlier.

Specificity, is used to measure whether a response is specific to a given context. For example, if a user says "I love Eurovision" and the model responds "Me too," then it would score 0% on specificity, since this response could be used in many different contexts. If it answers "Me too. I love Eurovision songs," then it would score 100%.

Interestingness, is used to measure if a dialogue is likely to "catch someone's attention" or "arouse their curiosity", or if it is unexpected, witty, or insightful.

Informativeness, measures the percentage of the dialogue that gives information. The information only needs to be correct within the dialogue (it is not fact-checked to external sources).

Groundedness, measures the percentage of the dialogue that contains claims about the external world that can be supported by external sources (it is fact-checked to external sources). So responses like "That's a great idea" that do not carry any external world information do not affect groundedness, but they do affect Informativeness. However, "Rafael Nadal is the winner of Roland Garros 2020" is an example of a grounded response.

The dialogue is great teaching material as is to teach the student to speak about the topic (on the given CEFR level) and use the exchange(s) (written in bold) in English. *

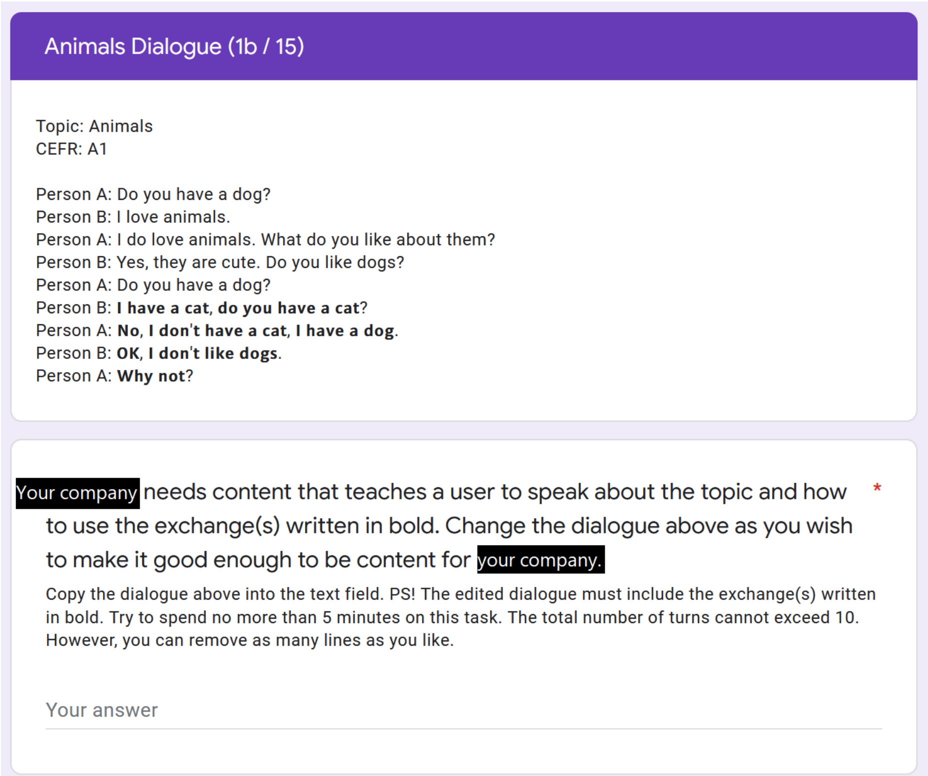
	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

Any other comments on why you agree or disagree to the previous statement?

Your answer

Figure 85: The continuation of the first task of the questionnaire, where the evaluation metrics in Figure 84 are described. The participant is also asked to give an evaluation of the teachability (in the middle box) of the extended dialogue on a 7-point Likert scale. The participant is also given the ability to give further comments (on the bottom) on how they agree or disagree with the teachability score they have given. After finishing these tasks, the participant is sent to the task shown in Figure 86.

A.3.3 Part 3: Make Adjustments to the Extended Dialogue


Animals Dialogue (1b / 15)

Topic: Animals
CEFR: A1

Person A: Do you have a dog?
Person B: I love animals.
Person A: I do love animals. What do you like about them?
Person B: Yes, they are cute. Do you like dogs?
Person A: Do you have a dog?
Person B: **I have a cat**, **do you have a cat**?
Person A: **No**, **I don't have a cat**, **I have a dog**.
Person B: **OK**, **I don't like dogs**.
Person A: **Why not**?

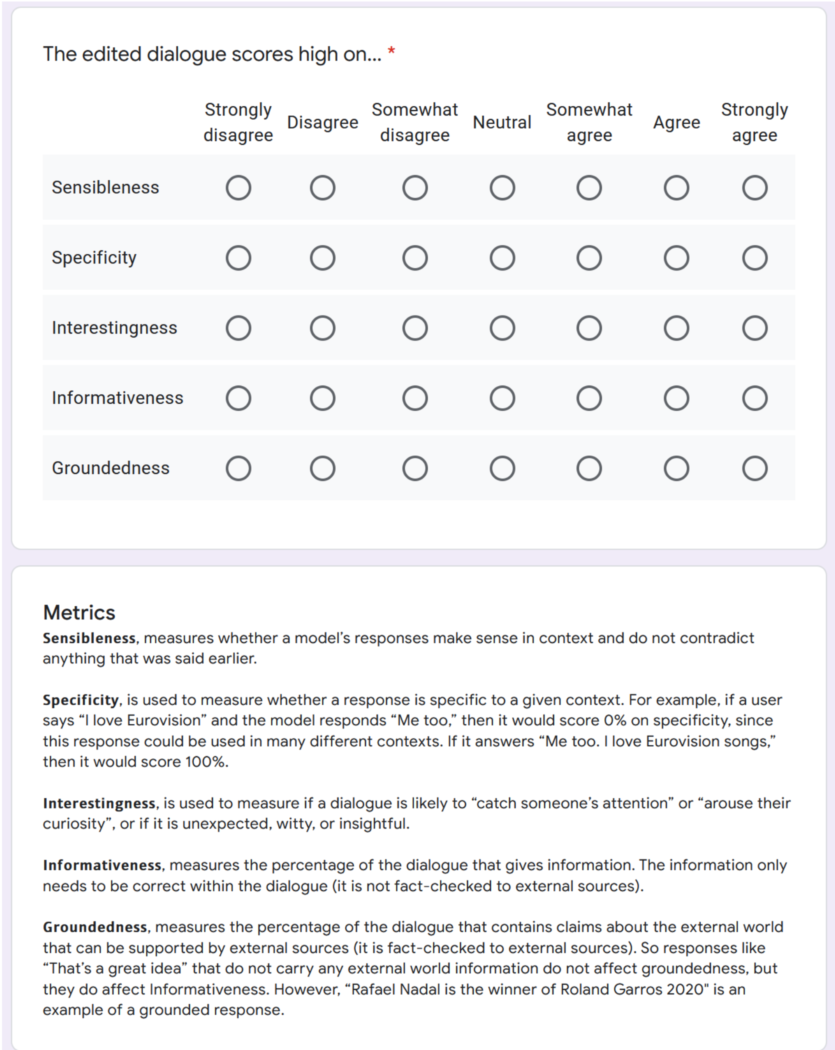
Your company needs content that teaches a user to speak about the topic and how to use the exchange(s) written in bold. Change the dialogue above as you wish to make it good enough to be content for your company. *

Copy the dialogue above into the text field. PS! The edited dialogue must include the exchange(s) written in bold. Try to spend no more than 5 minutes on this task. The total number of turns cannot exceed 10. However, you can remove as many lines as you like.

Your answer

Figure 86: The participant is first presented the extended dialogue again (from Figure 84). The task is to make adjustments to the extended dialogue, given the instructions (in the bottom box). After finishing the task, the participant is sent to the task shown in Figure 87. The company name is censored due to a non-disclosure agreement.

A.3.4 Part 4: Evaluate the Adjusted Dialogue



The edited dialogue scores high on... *

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
Sensibleness	○	○	○	○	○	○	○
Specificity	○	○	○	○	○	○	○
Interestingness	○	○	○	○	○	○	○
Informativeness	○	○	○	○	○	○	○
Groundedness	○	○	○	○	○	○	○

Metrics

Sensibleness, measures whether a model's responses make sense in context and do not contradict anything that was said earlier.

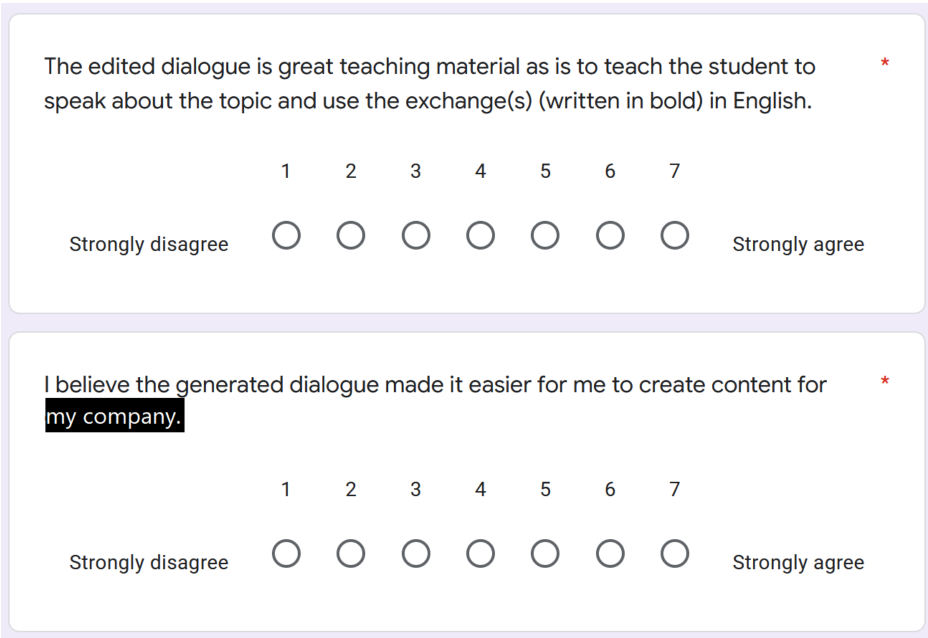
Specificity, is used to measure whether a response is specific to a given context. For example, if a user says "I love Eurovision" and the model responds "Me too," then it would score 0% on specificity, since this response could be used in many different contexts. If it answers "Me too. I love Eurovision songs," then it would score 100%.

Interestingness, is used to measure if a dialogue is likely to "catch someone's attention" or "arouse their curiosity", or if it is unexpected, witty, or insightful.

Informativeness, measures the percentage of the dialogue that gives information. The information only needs to be correct within the dialogue (it is not fact-checked to external sources).

Groundedness, measures the percentage of the dialogue that contains claims about the external world that can be supported by external sources (it is fact-checked to external sources). So responses like "That's a great idea" that do not carry any external world information do not affect groundedness, but they do affect Informativeness. However, "Rafael Nadal is the winner of Roland Garros 2020" is an example of a grounded response.

Figure 87: The participant is asked to evaluate the adjusted dialogue (from Figure 86) on the metrics explained in the bottom box. The next tasks are shown in Figure 88.


The edited dialogue is great teaching material as is to teach the student to speak about the topic and use the exchange(s) (written in bold) in English. *

	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

I believe the generated dialogue made it easier for me to create content for my company. *

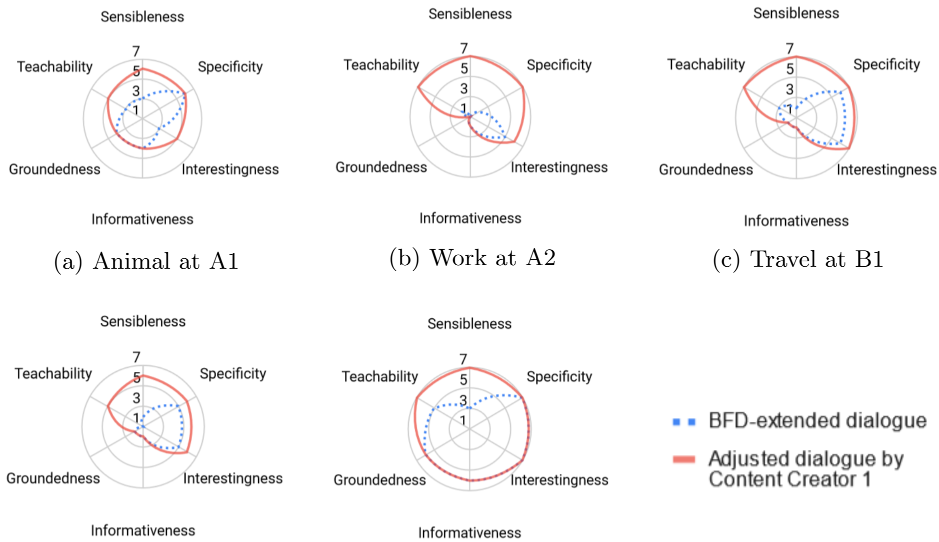
	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

Figure 88: The participant is asked to evaluate the adjusted dialogue (from Figure 87) on teachability (in the top box) and usefulness (in the bottom box). The company name is censored due to a non-disclosure agreement.

A.4 Experiment 3: User Study Responses

In this appendix, we show the results from the user study of Experiment 3 (see Section 5.4). The results are divided into an appendix for each of the ways the dialogues are extended: with the BFD Generator (see Appendix A.4.1), the BD Generator (see Appendix A.4.2), and the FD Generator (see Appendix A.4.3). In those appendices, the evaluations of extended dialogues compared to their adjusted counterpart are shown. In Appendix A.4.4, the evaluations, including the usefulness, of the extended dialogues using the BFD Generator, the BD Generator, and FD Generator are compared to each other. The actual extended dialogues and adjusted dialogues are given in Appendix A.4.5 (BFD-extended), Appendix A.4.6 (BD-extended), and Appendix A.4.7 (FD-extended).

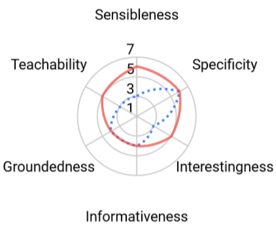
A.4.1 BFD Generator



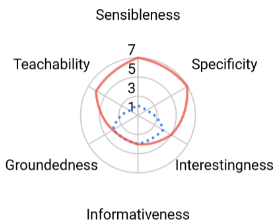
(a) Animal at A1 (b) Work at A2 (c) Travel at B1

(d) Interests at B2 (e) Politics at B2

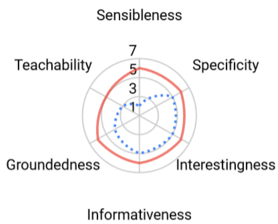
Figure 89: Evaluations of BFD-extended dialogues and self-evaluations of the adjusted dialogues by Content Creator 1. The topic and CEFR level of the dialogues are given below each radar chart.



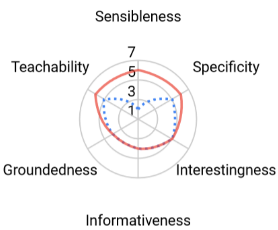
(a) Animal at A1



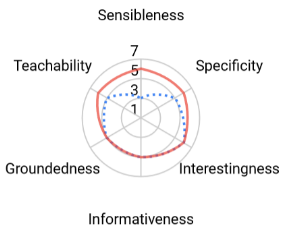
(b) Work at A2



(c) Travel at B1



(d) Interests at B2



(e) Politics at B2

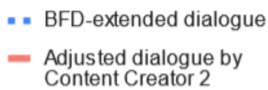
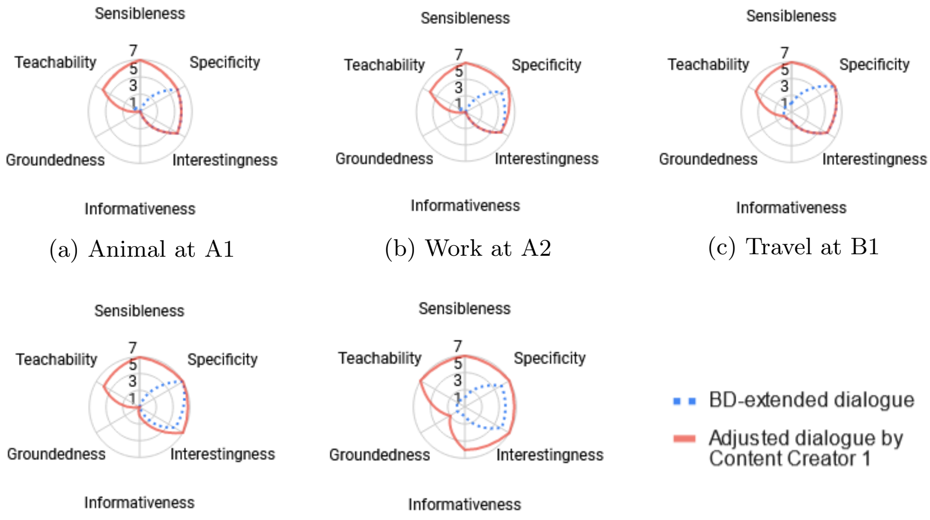


Figure 90: Evaluations of BFD-extended dialogues and self-evaluations of the adjusted dialogues by Content Creator 2. The topic and CEFR level of the dialogues are given below each radar chart.

A.4.2 BD Generator



(a) Animal at A1 (b) Work at A2 (c) Travel at B1

(d) Interests at B2 (e) Politics at B2

Figure 91: Evaluations of BD-extended dialogues and self-evaluations of the adjusted dialogues by Content Creator 1. The topic and CEFR level of the dialogues are given below each radar chart.

(a) Animal at A1

(b) Work at A2

(c) Travel at B1

(d) Interests at B2

(e) Politics at B2

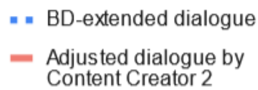
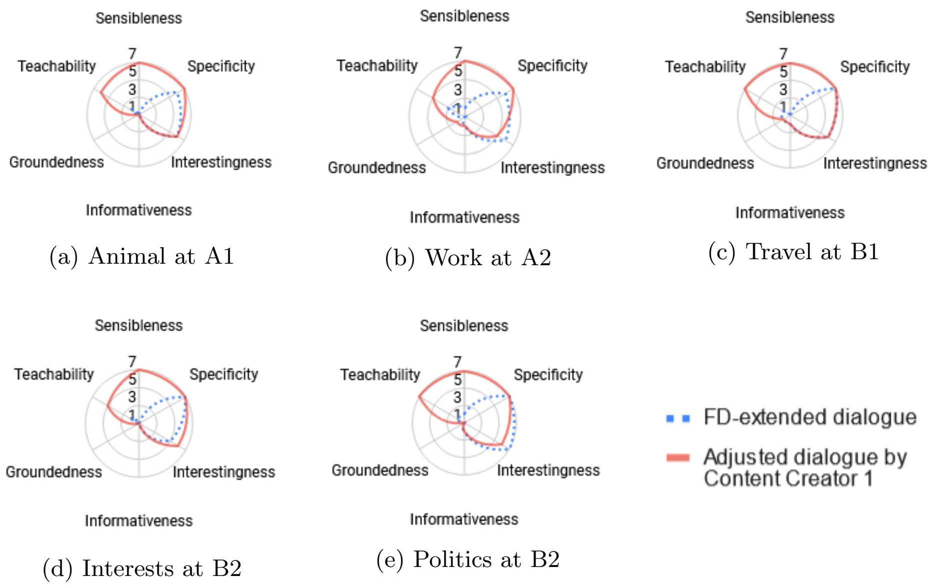


Figure 92: Evaluations of BD-extended dialogues and self-evaluations of the adjusted dialogues by Content Creator 2. The topic and CEFR level of the dialogues are given below each radar chart.

A.4.3 FD Generator

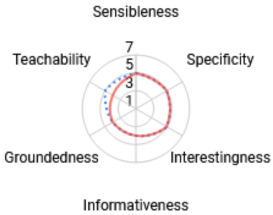


(a) Animal at A1 (b) Work at A2 (c) Travel at B1

(d) Interests at B2 (e) Politics at B2

Figure 93: Evaluations of FD-extended dialogues and self-evaluations of the adjusted dialogues by Content Creator 1. The topic and CEFR level of the dialogues are given below each radar chart.

(a) Animal at A1



(b) Work at A2

(c) Travel at B1

(d) Interests at B2

(e) Politics at B2

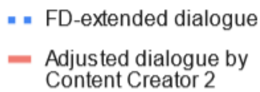
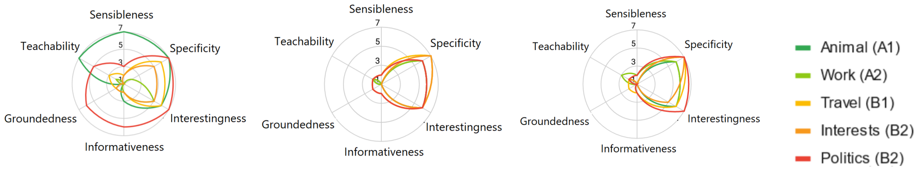


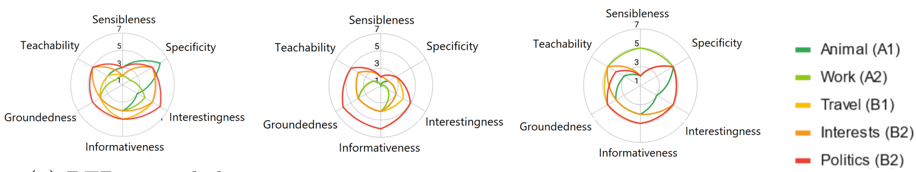
Figure 94: Evaluations of FD-extended dialogues and self-evaluations of the adjusted dialogues by Content Creator 2. The topic and CEFR level of the dialogues are given below each radar chart.

A.4.4 All Generators



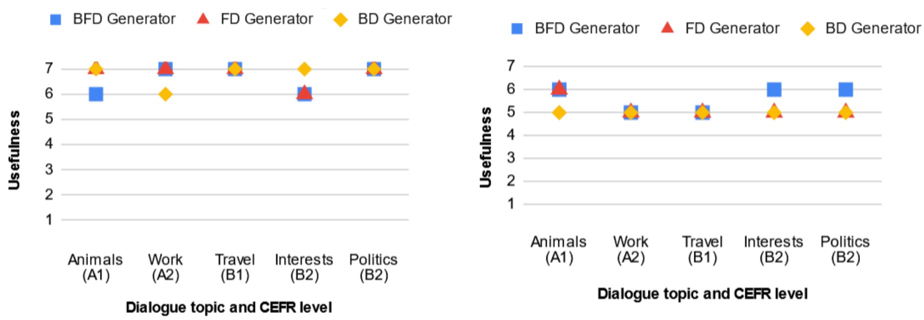
(a) BFD-extended (b) BD-extended (c) FD-extended

Figure 95: Evaluations of all BFD-, BD-, and FD-extended dialogues by Content Creator 1.



(a) BFD-extended (b) BD-extended (c) FD-extended

Figure 96: Evaluations of all BFD-, BD-, and FD-extended dialogues by Content Creator 2.



(a) Content Creator 1

(b) Content Creator 2

Figure 97: The feedback on the usefulness of extended dialogues for writing adjusted dialogues from Content Creator 1 (to the left) and Content Creator 2 (to the right). The horizontal axis represents the topic and CEFR level of the dialogues. The vertical axis represents the usefulness of the dialogues. The color and shape of the points indicate which generator was used to extend the dialogues (on the top).

A.4.5 BFD-extended and Adjusted Dialogues

BFD-extended Dialogue on the topic "Animal" and CEFR level A1
A: Do you have a dog?
B: I love animals.
A: I do love animals. What do you like about them?
B: Yes, they are cute. Do you like dogs?
A: Do you have a dog?
B: **I have a cat, do you have a cat?**
A: **No, I don't have a cat, I have a dog.**
B: **OK, I don't like dogs.**
A: **Why not?**
Content Creator 1's Adjusted Dialogue
A: I love animals. Do you?
B: Yes, I love animals.
A: What do you like about them?
B: I think all animals are cute. Do you like dogs?
A: Yes, I like dogs. Do you have a dog?
B: **I have a cat, do you have a cat?**
A: **No, I don't have a cat, I have a dog.**
B: **OK, I don't like dogs.**
A: **Why not?**
Content Creator 2's Adjusted Dialogue
A: Do you like animals?
B: I love animals.
A: I love animals too. What do you like about them?
B: They are cute. Do you like dogs?
A: Do you have a dog?
B: **I have a cat, do you have a cat?**
A: **No, I don't have a cat, I have a dog.**
B: **OK, I don't like dogs.**
A: **Why not?**

Table 1: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Animal", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BFD-extended Dialogue on the topic "Work" and CEFR level A2
A: I am a freelancer. I love it. What do you do?
B: I work in a company. What do you do?
A: I work in the field of human resources. What kind of work do you do?
B: **How is work?**
A: **It is great, I love my job!**
B: **What do you work with?**
A: **I work as a teacher. How about you?**
B: I work with kids. How do you teach them?
A: I teach them how to work. I also teach them how to drive and drive a car.
Content Creator 1's Adjusted Dialogue
A: I started in my new job last week.
B: Congratulations!
A: Thank you!
B: **How is work?**
A: **It is great, I love my job!**
B: **What do you work with?**
A: **I work as a teacher. How about you?**
B: I work as a lawyer. Which subjects do you teach?
A: I teach English and Maths.
Content Creator 2's Adjusted Dialogue
A: Hi, how are you?
B: I am good, How are you?
A: I am fine, thank you.
B: **How is work?**
A: **It is great, I love my job!**
B: **What do you work with?**
A: **I work as a teacher. How about you?**
B: I work as a teacher as well!
A: Really? Where do you work?

Table 2: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Work", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BFD-extended Dialogue on the topic “Travel” and CEFR level B1
A: I have! What is your favorite travel destination?
B: I love traveling. I love to travel, do you?
A: Yes, I’ve been to Greece. Have you been to Greece?
B: **What are your plans for the summer holiday?**
A: **We are going to Greece. I can’t wait!**
B: **That’s great, I love Greece! Have you been there before?**
A: **No, it’s my first time. Do you have any recommendations?**
B: I don’t really know much about travel, but I know that I love to travel.
A: I have heard that travel is one of the most important activities for the human race. Do you know if that is true?
Content Creator 1’s Adjusted Dialogue
A: I love to travel, do you?
B: Yes, I love traveling. I’m going to Italy this summer.
A: That sounds nice!
B: **What are your plans for the summer holiday?**
A: **We are going to Greece. I can’t wait!**
B: **That’s great, I love Greece! Have you been there before?**
A: **No, it’s my first time. Do you have any recommendations?**
B: I haven’t been in Greece, so I don’t know. Have you been in Italy and can give me some recommendations?
A: Yes, I have been in Italy. You should visit the Colosseum in Rome!
Content Creator 2’s Adjusted Dialogue
A: Do you like to travel?
B: I love traveling. I love to travel, do you?
A: Yes, I love to travel as well.
B: **What are your plans for the summer holiday?**
A: **We are going to Greece. I can’t wait!**
B: **That’s great, I love Greece! Have you been there before?**
A: **No, it’s my first time. Do you have any recommendations?**
B: I think the Parthenon in Athens is amazing.
A: Then we will definitely go there. Do you have any other recommendations?

Table 3: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic “Travel”, is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BFD-extended Dialogue on the topic "Interests" and CEFR level B2
A: Hello! Are you interested in gardening? I'm a gardener.
B: I love gardening.
A: Interesting! What do you like?
B: **I have just finished reading a book about agriculture**
A: **Was it any good?**
B: **Yes, I learned a lot from it?**
A: **I'm not that interested in agriculture.**
B: **Why not?**
A: I'm not sure but I would love to.
Content Creator 1's Adjusted Dialogue
A: Gardening is a passion of mine. Are you interested in gardening?
B: I'm quite interested in gardening, but sadly I don't have green fingers.
A: I can learn you some tricks of the trade, if you want! But what other interests do you have?
B: **I have just finished reading a book about agriculture**
A: **Was it any good?**
B: **Yes, I learned a lot from it?**
A: **I'm not that interested in agriculture.**
B: **Why not?**
Content Creator 2's Adjusted Dialogue
A: Hello! Are you interested in gardening? I'm a gardener.
B: I love gardening.
A: Interesting! What do you like?
B: **I have just finished reading a book about agriculture**
A: **Was it any good?**
B: **Yes, I learned a lot from it?**
A: **I'm not that interested in agriculture.**
B: **Why not?** B: I'm more interested in plants and flowers.

Table 4: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Interests", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BFD-extended Dialogue on the topic "Politics" and CEFR level B2
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**
A: I can, but I don't think I could convince anyone that climate change is the most important issue in the world today.
B: I guess so. I do like the climate, but it is not a great topic for a discussion.
A: That is true. It is a very serious topic. The climate is a major issue in the world today.
B: I agree. It is a very serious topic. I am thinking of the topic of the next election, maybe the last one.
A: Maybe the last one. I really want to go to the United Nations next year and see if I can convince people that climate change is the most important issue in the world today.
Content Creator 1's Adjusted Dialogue
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**
A: First I would like to hear why you don't agree, then I can try to convince you!
Content Creator 2's Adjusted Dialogue
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**
A: I can, if you are prepared to listen with an open mind?
B: What do you mean by that? I know everything there is to know about the weather.
A: I don't want to talk about this after all.
B: I knew it! The climate change is a hoax and you know it too!

Table 5: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Politics", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

A.4.6 BD-extended and Adjusted Dialogues

BD-extended Dialogue on the topic "Animal" and CEFR level A1
A: Hi, do you like animals?
B: Have you ever been to a zoo?
A: I love animals.
B: What are some other interesting facts about animals?
A: Yes, I have a dog.
B: **I have a cat, do you have a cat?**
A: **No, I don't have a cat, I have a dog.**
B: **OK, I don't like dogs.**
A: **Why not?**
Content Creator 1's Adjusted Dialogue
A: My main interest is animals. Do you like animals?
B: Yes, I like animals.
A: Nice! Do you have pets?
B: **I have a cat, do you have a cat?**
A: **No, I don't have a cat, I have a dog.**
B: **OK, I don't like dogs.**
A: **Why not?**
Content Creator 2's Adjusted Dialogue
A: Hi, do you like animals?
B: Yes, I love animals!
A: I love animals too!
B: Do you have any pets?
A: Yes, I have a dog.
B: **I have a cat, do you have a cat?**
A: **No, I don't have a cat, I have a dog.**
B: **OK, I don't like dogs.**
A: **Why not?**

Table 6: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Animal", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BD-extended Dialogue on the topic "Work" and CEFR level A2
A: Hi, I'm a teacher. Do you work in a company called a Salesforce?
B: Hi! Do you work as a teacher?
A: Yes. I work in a company called a Salesforce. What's your favorite job?
B: Do you like to work?
A: I work as a teacher. Do you like to work?
B: **How is work?**
A: **It is great, I love my job!**
B: **What do you work with?**
A: **I work as a teacher. How about you?**
Content Creator 1's Adjusted Dialogue
A: Nice to see you again!
B: Thanks, you too.
A: **How is work?**
B: **It is great, I love my job!**
A: **What do you work with?**
B: **I work as a teacher. How about you?**
Content Creator 2's Adjusted Dialogue
A: Hi, nice to meet you!
B: Hi! Nice to meet you too!
A: Do you come here often?
B: Yes, I try to come as often as possible. It's nice to get away from work.
A: I agree, sometimes it is nice.
B: **How is work?**
A: **It is great, I love my job!**
B: **What do you work with?**
A: **I work as a teacher. How about you?**

Table 7: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Work", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BD-extended Dialogue on the topic "Travel" and CEFR level B1
A: Hi! I love travel, what do you like about traveling?
B: Hi, I'm a travel enthusiast. Have you been traveling?
A: Hi! I love traveling, what are you doing?
B: Hi, I have been traveling a lot lately. Do you have a favorite place to go to?
A: I do. We are going to Italy. Do you have any plans?
B: **What are your plans for the summer holiday?**
A: **We are going to Greece. I can't wait!**
B: **That's great, I love Greece! Have you been there before?**
A: **No, it's my first time. Do you have any recommendations?**
Content Creator 1's Adjusted Dialogue
A: I can't wait for the summer holiday!
B: Me neither, I love to travel and this year I'm going to Spain.
A: That sounds great! Is this your first time going to Spain?
B: Yes, it's my first time, so I'm really excited.
A: I can see that!
B: **What are your plans for the summer holiday?**
A: **We are going to Greece. I can't wait!**
B: **That's great, I love Greece! Have you been there before?**
A: **No, it's my first time. Do you have any recommendations?**
Content Creator 2's Adjusted Dialogue
A: I'm really looking forward to my summer holiday!
B: **What are your plans for the summer holiday?**
A: **We are going to Greece. I can't wait!**
B: **That's great, I love Greece! Have you been there before?**
A: **No, it's my first time. Do you have any recommendations?**

Table 8: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Travel", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BD-extended Dialogue on the topic "Interests" and CEFR level B2
A: I have never read. What are you interested in?
B: I love reading. What is your favorite book?
A: I have not read it, do you?
B: I have not. What was the book about?
A: **I have just finished reading a book about agriculture**
B: **Was it any good?**
A: **Yes, I learned a lot from it?**
B: **I'm not that interested in agriculture.**
A: **Why not?**
Content Creator 1's Adjusted Dialogue
A: I love to read, but I don't have the time to read as much as I would want.
B: I agree, but I try to read some during the weekends and holidays at least.
A: Yes, me too. Reading is my favourite hobby.
B: Mine too! What was the last book you read?
A: **I have just finished reading a book about agriculture**
B: **Was it any good?**
A: **Yes, I learned a lot from it?**
B: **I'm not that interested in agriculture.**
A: **Why not?**
Content Creator 2's Adjusted Dialogue
A: I love reading. Books are great!
B: Have you read any good books lately?
A: **I have just finished reading a book about agriculture**
B: **Was it any good?**
A: **Yes, I learned a lot from it?**
B: **I'm not that interested in agriculture.**
A: **Why not?**

Table 9: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Interests", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

BD-extended Dialogue on the topic "Politics" and CEFR level B2
A: I am a Democrat. Do you vote for the Green Party?
B: I like to vote for the Green Party. Do you have a favorite politician?
A: Yes I do. Do you know anything about politics?
B: Yes. I am a Democrat. I have always been a Democrat. Do you know anything about politics?
A: Yes I voted for Jill Stein. Do you know anything about it?
B: **Did you vote in this year's election?**
A: **Yes, I did. Did you?**
B: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
A: **Really? I don't agree at all. Can you try to convince me?**
Content Creator 1's Adjusted Dialogue
A: I don't understand people who don't care about politics at all.
B: I agree, it's important to try to understand what the different parties stand for at least.
A: Definetely, if not you can't make a conscious choise in elections!
B: **Did you vote in this year's election?**
A: **Yes, I did. Did you?**
B: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
A: **Really? I don't agree at all. Can you try to convince me?**
Content Creator 2's Adjusted Dialogue
B: Are you interested in politics?
A: I'm not very interested.
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**

Table 10: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Politics", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

A.4.7 FD-extended and Adjusted Dialogues

FD-extended Dialogue on the topic "Animal" and CEFR level A1
A: **I have a cat, do you have a cat?**
B: **No, I don't have a cat, I have a dog.**
A: **OK, I don't like dogs.**
B: **Why not?**
A: I like cats, they are amazing. I like them too!
B: Yes, I have a dog, I like him too!
A: That's great! Do you have any favorite breed of dogs?
B: I like the dogs, they are so cute and cute.
A: I have a dog, I like them too.
Content Creator 1's Adjusted Dialogue
A: **I have a cat, do you have a cat?**
B: **No, I don't have a cat, I have a dog.**
A: **OK, I don't like dogs.**
B: **Why not?**
A: I'm afraid of them.
B: Oh, I see. Do you have any other pets?
A: Yes, I have a parrot too. Do you have any other pets?
Content Creator 2's Adjusted Dialogue
A: **I have a cat, do you have a cat?**
B: **No, I don't have a cat, I have a dog.**
A: **OK, I don't like dogs.**
B: **Why not?**
A: I like cats, better. They are less work.
B: I prefer dogs. Dogs are good company.
A: I think cats are good company too.
B: I think cats are arrogant. Dogs are more friendly.
A: I do not agree at all.

Table 11: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Animal", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

FD-extended Dialogue on the topic "Work" and CEFR level A2
A: **How is work?**
B: **It is great, I love my job!**
A: **What do you work with?**
B: **I work as a teacher. How about you?**
A: I work for a teacher. What kind of work do you do?
B: I work with a teacher to provide lessons to students. I do a lot of tutoring.
A: What kind of work do you do?
B: I do a lot of teaching. I teach all kinds of topics.
A: That is cool!
Content Creator 1's Adjusted Dialogue
A: **How is work?**
B: **It is great, I love my job!**
A: **What do you work with?**
B: **I work as a teacher. How about you?**
A: I work as a teacher's assistant. Which subjects do you teach?
B: I teach Norwegian as a second language. What do you do in your job as a teacher's assistant?
Content Creator 2's Adjusted Dialogue
A: **How is work?**
B: **It is great, I love my job!**
A: **What do you work with?**
B: **I work as a teacher. How about you?**
A: I work as a teacher as well!
B: Really? Where do you teach?
A: I teach at Senior Falls High School.
B: I also teach there! How come I haven't seen you before.
A: I do not know. This is strange.

Table 12: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Work", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

FD-extended Dialogue on the topic "Travel" and CEFR level B1
A: **What are your plans for the summer holiday?**
B: **We are going to Greece. I can't wait!**
A: **That's great, I love Greece! Have you been there before?**
B: **No, it's my first time. Do you have any recommendations?**
A: I love traveling to Greece, I have been there before. What do you like about traveling to Greece?
B: I like traveling to Greece, because it is a great country and I love to go on holidays.
A: I love going to Greece! I like how the city is so beautiful. What are some things you like about it?
B: I love it. The city is great. The Greek word "greek" means "southern" and "southern" means "west."
A: That's interesting! I have never been to Greece before. Do you know what the geography is like there?
Content Creator 1's Adjusted Dialogue
A: **What are your plans for the summer holiday?**
B: **We are going to Greece. I can't wait!**
A: **That's great, I love Greece! Have you been there before?**
B: **No, it's my first time. Do you have any recommendations?**
A: Well, I prefer just being on the beach all day, so I recommend that!
B: I see! I don't have the patience to lie in the sun all day, so I think I'll do a lot of sightseeing, if it doesn't get too hot. What was the temperature like when you were there?
A: It was in the summertime, so about 30 C, which I think is the perfect temperature.
B: 30 C is a bit too hot for me, but it will be nice nevertheless. What are your plans for the summer holiday?
Content Creator 2's Adjusted Dialogue
A: **What are your plans for the summer holiday?**
B: **We are going to Greece. I can't wait!**
A: **That's great, I love Greece! Have you been there before?**
B: **No, it's my first time. Do you have any recommendations?**
A: The islands are the best part. I don't recommend going to Athens.
B: Why is that?
A: It is noisy and crowded. On the islands you can relax and the people are more friendly there.
B: Then I will travel to the islands! I'm looking forward to it!

Table 13: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Travel", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

FD-extended Dialogue on the topic "Interests" and CEFR level B2
A: **I have just finished reading a book about agriculture**
B: **Was it any good?**
A: **Yes, I learned a lot from it?**
B: **I'm not that interested in agriculture.**
A: **Why not?**
B: I've read a lot about it, and I like it.
A: I didn't know that. I'm not sure if it's a good book or not.
B: I've never read it, what else can you tell me about it?
A: I'm not sure. I'm not sure if it's a good book or not.
Content Creator 1's Adjusted Dialogue
A: **I have just finished reading a book about agriculture**
B: **Was it any good?**
A: **Yes, I learned a lot from it?**
B: **I'm not that interested in agriculture.**
A: **Why not?**
B: Well, first of all I've lived in the city all my life so I have no experience with agriculture.
A: I see. Have you ever visited a farm or spent time in the countryside?
B: No, I haven't, and I don't want to either - I just love the city and all it's possibilites so much.
A: OK, so what are your favourite things about the city?
Content Creator 2's Adjusted Dialogue
A: **I have just finished reading a book about agriculture**
B: **Was it any good?**
A: **Yes, I learned a lot from it?**
B: **I'm not that interested in agriculture.**
A: **Why not?**
B: I'm more interested in flowers and plants.
A: I see. Can you tell me something interesting about flowers?
B: Did you know that the world's biggest flower smells like rotting flesh?
A: I didn't know. That sounds awful!

Table 14: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Interests", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

FD-extended Dialogue on the topic "Politics" and CEFR level B2
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**
A: I think it would be helpful if you were to go to the elections in the United Kingdom and vote. You would get to know the government.
B: I would love to. What country did you vote for?
A: I was in Scotland. I am from Scotland, and I am not sure what country voted in the election.
B: I did not vote in the referendum. What was the difference between the two?
A: I am not sure, but I think that the Green Party was the better candidate.
Content Creator 1's Adjusted Dialogue
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**
A: I sure can, I'm very passionate about these issues. Are you ready to hear me out?
B: Sure, go ahead.
Content Creator 2's Adjusted Dialogue
A: **Did you vote in this year's election?**
B: **Yes, I did. Did you?**
A: **Yes, I voted for the Green Party. I think global warming and environment protection is the most important matter in the world today.**
B: **Really? I don't agree at all. Can you try to convince me?**
A: I think you should start by reading news from varied sources.
B: What do you mean by that?
A: I can not convince you if you already have made up your mind.
B: That's what I thought.

Table 15: The extended dialogue (on the top), using the dialogue snippet (in bold) and the input topic "Politics", is turned into adjusted dialogues by Content Creator 1 (in the middle) and Content Creator 2 (on the bottom). The content creators consider the given CEFR level when adjusting the dialogues.

NTNU

Norwegian University of
Science and Technology