applied sciences

MDPI

Article

Automatic Detection of Inconsistencies and Hierarchical Topic Classification for Open-Domain Chatbots

Mario Rodríguez-Cantelar [1,†], Marcos Estecha-Garitagoitia [2,†], Luis Fernando D'Haro [2,*], Fernando Matía [1] and Ricardo Córdoba [2]

[1] Intelligent Control Group (ICG), Centre for Automation and Robotics (CAR) UPM-CSIC, Universidad Politécnica de Madrid, C. José Gutiérrez Abascal, 2, 28006 Madrid, Spain; mario.rcantelar@upm.es (M.R.-C.); fernando.matia@upm.es (F.M.)
[2] Speech Technology and Machine Learning Group (THAU), ETSI de Telecomunicación, Universidad Politécnica de Madrid, Av. Complutense, 30, 28040 Madrid, Spain; marcos.estecha.garitagoitia@upm.es (M.E.-G.); ricardo.cordoba@upm.es (R.C.)
* Correspondence: luisfernando.dharo@upm.es; Tel.: +34-9106-72174
† These authors contributed equally to this work.

Abstract: Current State-of-the-Art (SotA) chatbots are able to produce high-quality sentences, handling different conversation topics and larger interaction times. Unfortunately, the generated responses depend greatly on the data on which they have been trained, the specific dialogue history and current turn used for guiding the response, the internal decoding mechanisms, and ranking strategies, among others. Therefore, it may happen that for semantically similar questions asked by users, the chatbot may provide a different answer, which can be considered as a form of hallucination or producing confusion in long-term interactions. In this research paper, we propose a novel methodology consisting of two main phases: (a) hierarchical automatic detection of topics and subtopics in dialogue interactions using a zero-shot learning approach, and (b) detecting inconsistent answers using k-means and the Silhouette coefficient. To evaluate the efficacy of topic and subtopic detection, we use a subset of the DailyDialog dataset and real dialogue interactions gathered during the Alexa Socialbot Grand Challenge 5 (SGC5). The proposed approach enables the detection of up to 18 different topics and 102 subtopics. For the purpose of detecting inconsistencies, we manually generate multiple paraphrased questions and employ several pre-trained SotA chatbot models to generate responses. Our experimental results demonstrate a weighted F-1 value of 0.34 for topic detection, a weighted F-1 value of 0.78 for subtopic detection in DailyDialog, then 81% and 62% accuracy for topic and subtopic classification in SGC5, respectively. Finally, to predict the number of different responses, we obtained a mean squared error (MSE) of 3.4 when testing smaller generative models and 4.9 in recent large language models.

Keywords: chatbots; inconsistent responses; zero-shot topic detection; clustering

check for updates

Citation: Rodríguez-Cantelar, M.; Estecha-Garitagoitia, M.; D'Haro, L.F.; Matía, F.; Córdoba, R. Automatic Detection of Inconsistencies and Hierarchical Topic Classification for Open-Domain Chatbots. *Appl. Sci.* **2023**, *13*, 9055. https://doi.org/10.3390/app13169055

Academic Editor: Rocco Zaccagnino

Received: 18 June 2023
Revised: 28 July 2023
Accepted: 2 August 2023
Published: 8 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The extended usage of open-domain conversational systems has seen a significant growth in recent years, driven by a confluence of several factors. First, there is a growing interest among companies in offering alternative modes of communication with customers and potential users while simultaneously seeking avenues to streamline operational expenses. Additionally, users themselves have grown increasingly accustomed to such systems, expecting prompt responses to their queries or requests, as well as anticipating enhanced comprehension of their requirements and even entertainment. Lastly, advances in technology have led to notable enhancements in quality, consistency, knowledge, engagement, and even empathy within these conversational systems.

In terms of technology advancements, the utilization of pre-trained large language models (LLMs), coupled with information retrieval strategies and controlled generation methodologies, has led to the development of noteworthy chatbot implementations such as BlenderBot v3.0 [1], LAMDa [2], or GPT-4 [3]. However, it is worth noting, as we demonstrated with our chatbot Genuine^2 [4], developed for the Alexa Prize Socialbot Challenge (SGC4) [5,6], that chatbots can exhibit varying responses to semantically similar user queries due to several factors. These factors include: (a) the presence of contradictory information in the training data, (b) the influence of different dialogue histories and user question variations (i.e., paraphrases), and (c) the utilization of distinct decoding mechanisms for response generation (e.g., top-k [7], top-p [8], greedy search, beam search), which, over extended interactions, may give rise to instances of confusion or can be seen as a form of hallucination [9] in the chatbot.

On the other hand, open-domain chatbots face the challenge of engaging in multitopic conversations. Therefore, the precise and detailed detection of the different topics and subtopics that can occur within a dialogue assumes great significance in effectively managing the ongoing interaction, determining the subsequent states of the conversation, and providing pertinent information. For instance, consider a scenario where a user discusses the musical endeavors of their favorite singer. In such cases, the chatbot must adeptly discern the high-level topic (e.g., entertainment) and corresponding subtopics (e.g., music and singer) being discussed by the user. This capability is crucial in order to avoid generating responses that might disrupt the flow of the conversation. Regrettably, most existing topic classifiers exhibit static behavior; that is, they are limited to classifying only the predetermined set of topic labels seen during the training phase. Therefore, they are not able to handle new topics or subtopics.

In this research paper, we present practical solutions to address the aforementioned limitations. First, we outline a hierarchical algorithm that takes advantage of zero-shot learning approaches to allow a hierarchical classification of topics and subtopics. The efficacy of the proposed algorithm is evaluated using a sample of 1000 dialogue turns, randomly selected and annotated from the DailyDialog dataset, covering up to 18 distinct topics and 102 subtopics (Table 1). Furthermore, we tested this approach in real human–chatbot conversations obtained during our participation in the Alexa Prize Socialbot Challenge (SGC5) competition (https://www.amazon.science/alexa-prize/socialbot-grand-challenge/2022 (accessed on 14 July 2023)). Second, we introduce an automated algorithm designed to identify inconsistencies or different semantic variabilities in chatbot responses. This algorithm uses clustering strategies to identify multiple semantic variations in chatbot responses to similar users' questions, therefore detecting the cases where the provided answers exhibit disparities (inconsistent responses).

Table 1. List of proposed topics to predict.

Proposed Topics List	
Topics	**Subtopics**
animals	cats, dogs, pets.
art	ballet, cinema, museum, painting, theater.
books	author, genre, Harry Potter, plot, title.
education	history, mark, professor, school, subject, university.
family	parents, friends, relatives, marriage, children.
fashion	catwalk, clothes, design, dress, footwear, jewel, model.
finance	benefits, bitcoins, buy, finances, investment, sell, stock market, taxes.
food	drinks, fish, healthy, meal, meat, vegetables, dessert.
movies	actor, director, genre, plot, synopsis, title.
music	band, dance, genre, lyrics, rhythm, singer, song.
news	exclusive, fake news, interview, press, trending.
photography	camera, lens, light, optics, zoom.

Table 1. *Cont.*

Proposed Topics List	
Topics	**Subtopics**
science	math, nature, physics, robots, space.
sports	baseball, basketball, coach, exercise, football, player, soccer, tennis.
vehicle	bike, boat, car, failure, fuel, parts, plane, public transport, speed.
video games	arcade, computer, console, Nintendo, PlayStation, Xbox.
weather	cloudy, cold, hot, raining, sunny.
other	

The structure of this paper is as follows. Section 2 presents the State-of-the-Art on inconsistent responses and zero-shot classification. Section 3 delves into the comprehensive description of the proposed algorithms. Section 4 provides an elaborate account of the dataset used for the task of determining topics and subtopics, along with the process involved in designing the dataset used for the automatic detection of inconsistencies. Then, Section 5 describes the results obtained for both algorithms. Finally, Section 6 concludes this article by offering key insights and presenting avenues for future research.

2. Related Work

Automatic text classification for multiple labels representing different types or classes is a well-known problem. A variety of models and datasets have been developed with a wide range of performances [10]. Typically, models such as BERT [11], RoBERTa [12], or XLM-RoBERTa [13] have been used for this task.

Two of the most promising methods for fast classification and scalability are few-shot and zero-shot classification techniques. This type of classification allows the use of labels never seen during training by taking advantage of the semantic information of the selected names for the labels [14,15]. In few-shot, we provide a limited number of examples of the labels we want to classify. For zero-shot approaches, no examples are provided at all. These techniques allow for the use of new general labels. Thus, you can dynamically update the number of classes to classify [16].

On the other hand, inconsistency in the responses of conversational agents is a problem that even the best current chatbots still have. The ChatGPT model provides fun, information, advice, and influences users' moral judgment, but is sometimes inconsistent [17]. One of the reasons for these inconsistencies is that the models have been trained on large datasets containing similar topical information but with large variabilities in their content, which is good for robust and diverse generation, but potentially makes the model generate wrong or inconsistent responses [18]. Knowledge-based conversational models have been found to not only produce inconsistencies, but also to amplify them due to the quality of the training data [19].

Coherence is defined as the ability of a system to maintain a good conversational flow throughout a dialogue [20]. This is an essential property to pursue in an open-domain dialogue system that intends to converse with humans; however, it is a complex field of research due to the open nature of the dialogue. Automatic metrics such as BLEU [21] assess the degree of word overlap between a dialogue response and its corresponding gold response, but do not take into account sentence semantics [22], whereby they correlate weakly with human judgments, resulting in bias. To overcome this problem, metrics based on learning models such as ADEM [23], RUBER [24], and BERTRUBER [25] have been developed. There are other types of evaluation metrics, such as the GRADE [26] model based on graphical reasoning, which through graphical representations is able to extract thematic transitions throughout a dialogue and include them in the evaluation.

Sadly, evaluating inconsistencies in open-domain dialogue systems is challenging due to the problem of distinguishing between correct diverse answers and inconsistent ones. Researchers often rely on expensive and non-scalable human judgment experiments for response quality assessment [27]. One possible solution is training models using contrastive learning, optimizing implicit knowledge retrieval [28], but again it is challenging to create high-quality contrastive examples.

This work is based on the architecture and experiments described in [29]. However, this introduces multiple extensions in terms of the number of topics (from 13 to 18, around 40% more classes) and subtopics, datasets, and the need to perform new tests on the latest State-of-the-Art models based on LLMs. Special effort has been put into robustly automating the architecture and validating it with manually annotated data.

To further evaluate the performance of the inconsistency detection algorithm, the results have been extended using newer and more recent conversational agents. For this purpose, new topic-based questions and their respective paraphrases have been automatically generated with GPT-4 and manually annotated by experts. To test the reliability of the proposed algorithm, a new type of embedding encoder is introduced to confirm the findings of the original study, in which the proposed methodology is robust to variations in the embeddings used.

3. Architecture

In this section, we discuss the details of the algorithms proposed to address two key aspects in open-domain dialogue systems: hierarchical topic classification using Zero-Shot approaches and the automatic estimation of inconsistent responses.

3.1. Zero-Shot Topic Classification

The objective of this algorithm is to identify topics and subtopic classes present in the various turns of a dialogue in a scalable and adaptable manner, reducing or avoiding the need to train a specialized topic classifier. One of the key benefits of this approach is its ability to accommodate the dynamic expansion of topics or subtopics without requiring labeled data or retraining of the model.

Our algorithm is built on the zero-shot classification methodology [30], where classification is approached as an outcome of the Natural Language Inference (NLI) process. In our implementation, we used models specifically trained on NLI, available through the HuggingFace library.

More concretely, the proposed methodology described in [30] employs a pre-trained sequence pair classifier based on the Multi-Genre Natural Language Inference (MNLI) task. This approach involves mapping sequences into a shared latent space and evaluating the distance between them. Specifically, the NLI framework assumes the existence of two sentences: a “premise” and a “hypothesis”. The pre-trained model is then tasked with determining the validity of the hypothesis (i.e., whether it is an entailment or a contradiction) given the premise. In the context of our task, the “premise” corresponds to the sentence that requires the topic or subtopic classification, while each candidate label (for topic or subtopic embedded in a predefined prompt template) is treated as an independent “hypothesis”. The NLI model’s probability estimates to what extent the premise “entails” the different hypothesis (topics or subtopics) and uses it to rank the provided set of labels, ultimately selecting the label with the highest probability as the identified topic or subtopic.

Unfortunately, most topic classifiers are based on a set of fixed labels found during training. Therefore, they encounter scalability issues when confronted with a large number of labels and fail to identify subtopics, thereby impeding fine-grained analysis of conversational topics. Consequently, we decomposed the process into two consecutive or hierarchical steps. Initially, we establish a comprehensive set of words, denoted as topics, that encapsulate broad thematic categories such as entertainment, sports, family, science, music, etc. Subsequently, the Zero-Shot NLI classifier is employed on this set of labels, and the keyword exhibiting the highest probability (above a predetermined threshold) serves as the identified topic. Subsequently, this selected keyword proceeds to the second phase, where a more specific set of word labels (depending on the previously detected topic) is employed to detect subtopics, employing the same NLI model. This way, it is possible to handle a hierarchy of topics and subtopics while speeding up the inference process, avoiding having to search over the whole set of topic and subtopic words.

For example, within the primary category sports, the second set of labels could encompass baseball, basketball, soccer, or tennis, facilitating a more detailed identification of the subtopic of the conversation. This hierarchical approach grants several advantages: (a) the word set at each level can be tailored according to the specific domains and the desired number of topics/subtopics, and (b) the hierarchical structure can be expanded to accommodate additional sublevels of granularity, such as player, coach, stadium, team, and so forth. Consequently, this process allows for scalability and adaptability, as the expansion or adaptation of the terms can be easily made.

Below are the topics and subtopics that we used to classify each dialogue turn. This set of topics was created and improved by manual annotations carried out in [29] by annotators when a given sentence was assigned the "other" topic when classifying a turn. As a result, we went from being able to detect the original 13 topics (animals, books, cars, family, fashion, finance, food, movies, music, photography, sports, weather, and other) to detecting 18 topics and 102 subtopics (Table 1). There is an average of 6 subtopics for each of these 18 topics into which the turn can be classified at a more specific level of granularity. In our case, the list of topics and subtopics was manually produced; however, a more scalable method can be achieved by using Latent Diricleht Allocation (LDA, [31]) or BERTopic [32], also allowing the identification of relevant words that can be used as a topic or subtopic.

3.2. Detection of Response Inconsistencies

The second algorithm presented in this study focuses on automating the identification of response inconsistencies in chatbots. These inconsistencies can be attributed to several factors: (a) the presence of contradictory information within the training data of the language models, (b) the utilization of diverse decoding strategies and parameter configurations to generate response variations, such as top-p, top-k, greedy beam search, temperature, and others, (c) variations in the phrasing of questions posed by the user, (d) disparities in the dialogue context employed for generating the chatbot response, (e) the implementation of distinct re-ranking strategies for selecting the ultimate answers, and (f) the tendency in chatbots to agree with users (for instance, if the favorite user's color is blue, it is highly probable that the chatbot will also like blue, while later it could mention liking green if the user also likes this color), therefore showing some inconsistent preferences or persona profiles.

The proposed algorithm encompasses a sequence of five distinct steps, as illustrated in Figure 1. The initial step involves obtaining a collection of question–answer pairs, which can be acquired either from the chatbot's logs or from a manually constructed dataset comprising questions, paraphrases, and the corresponding chatbot responses. Details regarding the acquisition process are discussed in Section 4.

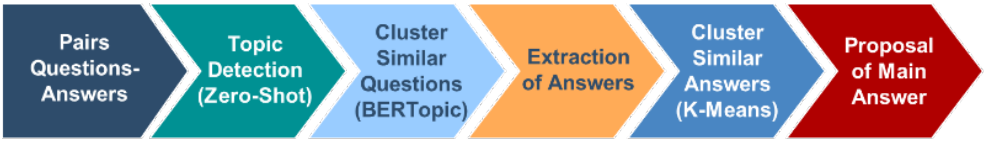

Pairs Questions-Answers
Topic Detection (Zero-Shot)
Cluster Similar Questions (BERTopic)
Extraction of Answers
Cluster Similar Answers (K-Means)
Proposal of Main Answer

Figure 1. Process flow for detection of response inconsistencies.

In the subsequent step, the classification of the topic for each question is performed using the Zero-Shot method outlined in Section 3.1. Specifically, questions that share the same topic or subtopic are identified and topic-related batches are generated to streamline and focus subsequent processes. In this research, ground-truth labels were utilized for this purpose.

For the third phase, we employed BERTopic [32] to cluster questions that exhibit similarity within the same topic or subtopic. This unsupervised algorithm aims to identify cohesive groups by extracting sentence vector embeddings using sentence-BERT [33].

Subsequently, the reduction in dimensionality of sentence embeddings is achieved using UMAP [34], followed by clustering with the HDBScan algorithm [35]. To extract the representation of each cluster, we utilize a variant of the TF-IDF formulation known as c-TF-IDF (class-based term frequency, inverse document frequency), which models the significance of words within each identified group. This formulation enables the detection of the most representative sentences for each cluster. Here, we can consider that each cluster consists of a representative question (i.e., centroid) and a whole set of semantically similar questions (i.e., paraphrases) where all answers associated to the questions in the same cluster are candidates of inconsistent answers to be detected in step four.

In the fourth phase, we collect all the answers associated with the clustered questions from the previous step. At this stage, and to reduce noise, we compute the cosine similarity between each answer and the centroid question. Answers with similarity below a specified threshold (in our case, 0.5) are then eliminated to further mitigate the possible noise during the question clustering and in large variabilities in the responses.

In the fifth step, we introduce a proxy mechanism to assess the presence of inconsistent responses. Here, we extract the sentence embeddings for each answer within a given question group. Subsequently, we apply the k-means algorithm to cluster all the answers, adjusting the number of clusters from two to the total number of answers minus one ($N-1$). We then calculate the Silhouette coefficient for each clustering scenario and the overall average Silhouette coefficient. The number of clusters that produce the highest number of samples with a Silhouette coefficient greater than the average value is selected as an approximation of the number of distinct chatbot responses generated.

$$\arg\max_i [Silhouette\ coef[i] > \frac{1}{N-1}\sum_{i=1}^{N-1}(Silhouette\ coef[i])] \quad (1)$$

If the system does not identify suitable clusters, a single cluster is considered (this is due to limitations in the k-means and Silhouette coefficient estimations). Furthermore, the system determines the centroid and identifies the closest answer to it, designating it as the optimal candidate for controlled generation (e.g., for utilization in the persona profile as used in TransferTransfo [36] or for constrained beam search generation [37]).

4. Data Annotations

In this section, we describe the data used to evaluate both the Zero-Shot hierarchical topic classification (Section 4.1) and the detection of inconsistent responses (Section 4.2).

4.1. Zero-Shot Topic Classification

To perform Zero-Shot topic detection in [29], the DailyDialog corpus [38] was used as an annotated data source. This dataset comprises human–human dialogues, encompassing a broad spectrum of common topics such as relationships, everyday life, and work-related issues. The primary purpose of these conversations is to exchange information and strengthen social connections. We processed approximately ∼13 k conversational turns and, through random selection, obtained a subset of 1000 turns for automated topic detection. For this purpose, we followed the methodology described in Section 3.1.

Afterward, a team of four expert evaluators conducted manual annotation on the DailyDialog subset, each evaluator handling 250 dialogues. They classified the dialogues into 13 different categories for high-level topic classification. As in [29], among the topics to choose from, one category was "other". However, annotators had the option to assign a more specific category for future research.

For this paper, we reuse the same DailyDialog subset collected in [29]. Here, four experts annotated 250 dialogues each. However, this time, we introduced new annotations at topic and subtopic level using the list indicated in Table 1.

To extend our analysis and test the robustness of our proposed approach, we included a second larger dataset from data collected by our Thaurus bot during our participation in the Alexa Prize Socialbot Challenge (SGC5) (https://www.amazon.science/alexa-prize/

teams/thaurus-2022 (accessed on 14 July 2023)). This dataset consists of conversations between users and our chatbot using echo devices. Unlike the conventional goal-oriented conversations provided by Alexa, in this competition, the conversations are open domain covering a long set of topics including movies, politics, news, health, arts, animals, etc. For these experiments, we randomly sampled dialogues collected during our participation in the competition.

Since this dataset is much larger than DailyDialog and we already had topic annotations at turn-level thanks to a pre-trained topic classifier included in the Amazon CoBot toolkit and provided to the participants [39], we decided to consider them as silver annotation labels (Unsupervised extracted labels obtained using a pre-trained model. Since this is an automatic process, there is no guarantee that all labels are correct, but if the pre-trained model is good, then it is expected that the labels are a good approximation to the case of manually labeling the data.) at topic level. The CoBot topic classifier returns the 11 topics shown in Table 2.

Table 2. List of labels predicted by the CoBot topic classifier.

CoBot Topics List			
1. books	2. general	3. interactive	4. inappropriate
5. movies	6. music	7. phatic	8. politics
9. science	10. sports	11. other	

As can be seen, this list overlaps in six topics with the list in Section 1.

Model Selection: In order to determine the quality of the silver labels provided by the CoBot topic classifier, but more specifically to classify good subtopics later on, we carried out experiments with several zero-shot models. Here, the goal was to select the model that could provide the highest correlation with the CoBot topic classifier. The hypothesis is that a high correlation with this pre-trained classifier will also be the best option to perform subtopic classification.

In this case, we tested multiple State-of-the-Art models based on DeBERTa [40], BART [41], and DistilBERT [42]. The specific models were: cross-encoder/nli-deberta-v3-base (https://huggingface.co/cross-encoder/nli-deberta-v3-base (accessed on 14 July 2023)), cross-encoder/nli-deberta-v3-large (https://huggingface.co/cross-encoder/nli-deberta-v3-large (accessed on 14 July 2023)), facebook/bart-large-mnli (https://huggingface.co/facebook/bart-large-mnli (accessed on 14 July 2023)), and typeform/distilbert-base-uncased-mnli (https://huggingface.co/typeform/distilbert-base-uncased-mnli (accessed on 14 July 2023)), all of which are available in the HuggingFace library. The results of these four models are presented in Section 5.1.

Then, once the best Zero-Shot model was selected, we used it to annotate at subtopic level and performed manual verification on a subset of the available data. Results for this task are shown in Sections 5.1.2 and 5.1.3.

4.2. Detection of Response Inconsistencies

In [29], to identify inconsistencies (see Section 3.2), we manually created a set of 15 unique canonical questions along with their corresponding paraphrases. We made sure that the paraphrases retained the same semantic meaning and intent. Table 3 provides several examples of the questions and the generated paraphrases. In total, we collected 109 questions, resulting in an average of 6.3 paraphrases per canonical question.

Then, we used four generative pre-trained chatbots to obtain their responses to the set of canonical questions and paraphrases. We selected: (a) DialoGPT-large [43] accessible via HuggingFace (https://huggingface.co/microsoft/DialoGPT-large (accessed on 14 July 2023)), (b) BlenderBot v2.0 (400M), (c) BlenderBot v2.0 (2.7B) [44–46], and (d) Seeker [47] without using its retrieval search module, the three of them available in ParlAI (https://parl.ai/projects/blenderbot2/ (accessed on 14 July 2023), https://parl.ai/projects/recipes/ (accessed on 14 July 2023), https://parl.ai/projects/seeker/ (accessed on 14 July 2023)).

Following that, four expert annotators reviewed each chatbot's answers for the canonical questions and paraphrases. The annotators' task was to count the number of distinct semantic responses generated by each chatbot. Ideally, each chatbot would produce a single coherent response independently of the variability of the paraphrases. However, our findings revealed an average of four distinct answers per chatbot.

Table 3. Example of original questions and paraphrases handcrafted for the detection of inconsistencies.

Question	Paraphrases
What is your favorite sport?	Which is the sport you like the most? My favorite sport is basketball, and yours? What kind of sport do you like?
What is your favorite book?	What is the title of your favorite book? Which book you always like to read? Hi!! I like reading, which book is your favorite one?
What is your job?	What do you do for a living? What do you do for work?

Then, to test the performance and robustness of our proposed algorithm with more recent LLMs and more challenging paraphrases, we decided to use GPT-4 (https://learn.microsoft.com/en-us/azure/cognitive-services/openai/ (accessed on 14 July 2023)) [3]. In this case, we asked GPT-4 to automatically generate a set of 15 original questions, and 7 paraphrases each (a total of 120 unique canonical questions), that a human could ask another one when they interact for the first time and want to get to know each other (e.g., when dating). These sentences and paraphrases were manually inspected and adapted if needed. Examples of the new generated canonical questions and paraphrases are shown in Table 4.

Finally, four State-of-the-Art pre-trained large language models were used to evaluate this new set of data. The new models tested were ChatGLM (6B) [48], BlenderBot3 (3B) [49], and GPT-4. Then, the chatbot responses were annotated in the same way as in the previous case, but this time only three experts participated. An overall improvement has been achieved with the new dialogue generation systems of up to 3% and 6% compared to the original results. These results are further discussed in Section 5.2, where the average number of different responses obtained for each new dialogue model using two different embedding encoders is shown.

Table 4. Example of new questions and paraphrases created with GPT-4 for the detection of inconsistencies.

Question	Paraphrases
What is your favorite hobby?	What leisure activity do you enjoy the most? Which pastime brings you the most satisfaction? What is the hobby that you find most appealing?
Who is your favorite superhero?	What superhero character do you admire the most? What is your preferred superhero of all time? Which fictional hero holds a special place in your heart?
What's your favorite type of cuisine?	Which cuisine do you find most appealing? Which type of cooking brings you the greatest delight? What style of food do you consider your favorite?

5. Results

5.1. Zero-Shot Topic Classification

This section describes the results obtained by testing our proposed methodology on hierarchical topic classification by using Zero-Shot approaches on two different datasets: DailyDialog and Alexa SGC5. First, we present our original results in DailyDialog only at the topic level (Section 5.1.1). Then, the selection of a robust Zero-Shot model on the Alexa SGC5 dataset and results at the topic and subtopic levels are presented (Section 5.1.2), and, finally, new results on DailyDialog at the subtopic level (Section 5.1.3).

5.1.1. Original Results on DailyDialog—Topic Level

First, we provide the results for zero-shot topic detection when using the original 13 high-level topics tested on 1000 random sentences from DailyDialog and using the original nli-deberta-base model (https://huggingface.co/cross-encoder/nli-deberta-base (accessed on 14 July 2023)). Figure 2 shows the F1 results for each topic (including the "other" class), and considering different thresholds for the Zero-Shot selection. As we can see, finding a global threshold is difficult since there are large variabilities depending on the topic; however, a good threshold will be between 0.8 and 0.9.

Figure 3 shows the F1 result when establishing a global threshold of 0.9. In this case, the macro-F1 score is 0.575, while the weighted F1 score is measured at 0.658. These results can be considered good in light of the number of topics considered and the fact that the model was not explicitly trained for the purpose of topic detection.

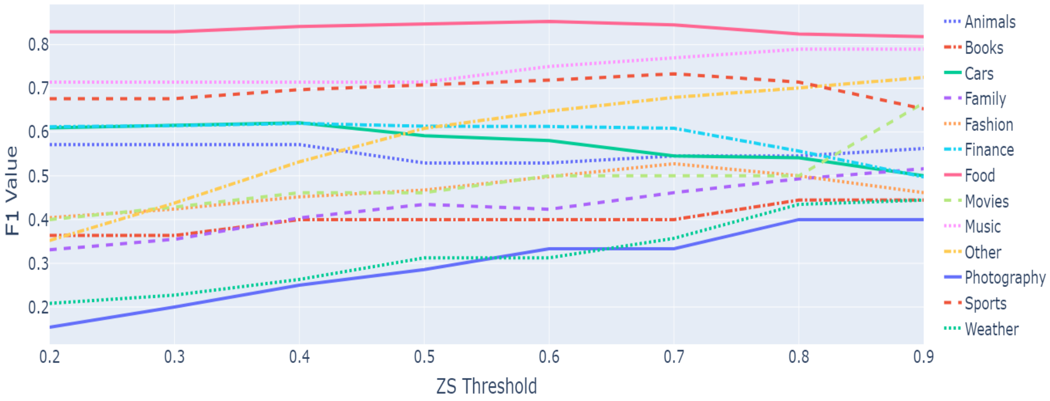


Figure 2. F1 variations for zero-shot classification over different thresholds and topics.

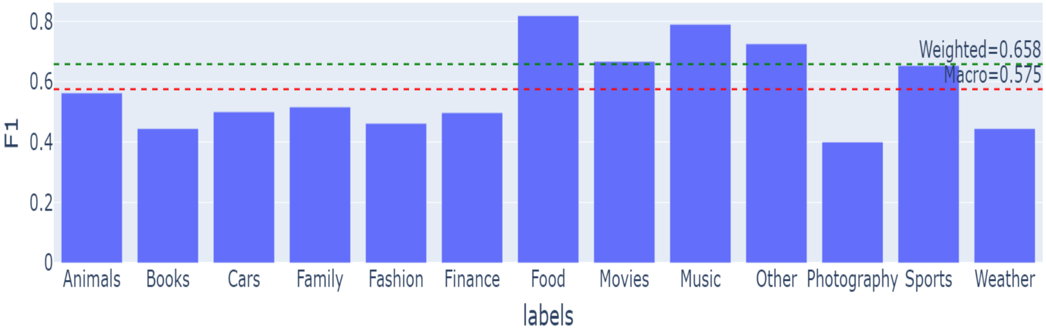


Figure 3. F1 results over different topics for th = 0.9. The macro-F1 is 0.575, while the weighted average F1 mean is 0.658.

The analysis of the figure reveals that the categories that show the most favorable outcomes include food, movies, music, and sports. This can be attributed to the careful selection of keywords employed in formulating the "hypothesis" sentence within the Natural Language Inference (NLI) framework, as well as possible similarities between the test data and the training data used for the selected NLI model. Conversely, categories such as books, photography, and weather pose challenges in terms of accurate recognition. Upon a deep review of the results, it was identified that some errors were due to the fact that the "premise" sentences lacked sufficient context for the model to accurately infer the topic, although humans might find it marginally easier to discern. For example, the sentence "Four by six, except this one. I want a ten-by-thirteen print of this one. Okay, they will be ready for you in an hour." is annotated as belonging to the category "photography", but for the model, it was considered books; in another case, the sentence "There are various magazines in the rack. Give me the latest issue of 'National Geographic'." was annotated as pertaining to the category "books", while the model classified it as "other".

5.1.2. New Results on SGC5—Topic and Subtopic Level

BART vs CoBot topic classifiers: As commented in Section 4.1, the SGC5 data were automatically labeled (silver labels) using a sample pre-trained topic classifier provided to the SGC5 teams as part of the CoBot architecture and then labeled with four NLI-based models (Deberta-v3, Deberta-v3-large, Bart-large, and Distillbert-uncased). These four models were tested to produce the same 11 topics provided by the CoBot classifier (Table 2). They also received the same amount of dialogue context, i.e., the current turn of the dialogue and the last three pairs of turns in the history. The goal here is to detect which of the four untrained Zero-Shot models provides the closest results when using the silver labels and then select that model for the subtopic level classification. Table 5 shows the results of the four models on the silver labels at topic level.

Table 5. NLI topic classifier models vs. CoBot topic classifier. The best F1-Score result among all models is shown underlined.

Model	Accuracy (%)	F1 Score	Precision	Recall
cross-encoder/nli-deberta-v3-base	10.08	0.08	0.22	0.15
cross-encoder/nli-deberta-v3-large	7.55	0.06	0.18	0.15
facebook/bart-large-mnli	25.98	<u>0.26</u>	0.30	0.29
typeform/distilbert-base-uncased-mnli	0.83	0.01	0.09	0.02

The model with the best performance was BART-large-mnli with an F1 Score of 0.26%; therefore, this model was selected to be used for the rest of the experiments. The confusion matrix between the CoBot topic classifier and BART is shown in Figure 4.

The confusion matrix shows that there is a high coincidence for sports, politics, movies, books, and music topics, while there is less coincidence for science, general, phatic, interactive, inappropriate, and other topics. This is probably due to the overlap in the data used to train the MNLI and the one used for testing. Moreover, topics with high matching are generally easier to classify, while the low matching topics are too open or prone to more subjective annotations.

Topic and Subtopic detection: To validate the possibility of performing automatic detection of topics and subtopics using a Zero-Shot approach (Section 3.1), the SGC5 data were automatically annotated at the topic and subtopic level using the BART model. In this case, the number of topics for BART were the 18 proposed as an improvement (Table 1), while for the CoBot topic classifier, they were the same 11 that the classifier was able to identify (Table 2).

The common topics between the BART and CoBot topic classifiers are: sports, movies, books, music, science, and other. The unique topics for the BART classifier are: animals, video games, politics, art, teacher, family, finance, cars, astronomy, food, photography, newspaper, fashion, climate conditions, and Facebook. As for the different topics that the CoBot topic classifier is able to detect, they are: general, phatic, interactive, and inappropriate.

To evaluate the real quality of automatic classifications performed using BART compared to those provided by the CoBot topic classifier, three experts manually evaluated 200 dialogue turns. These annotations were divided into two levels.

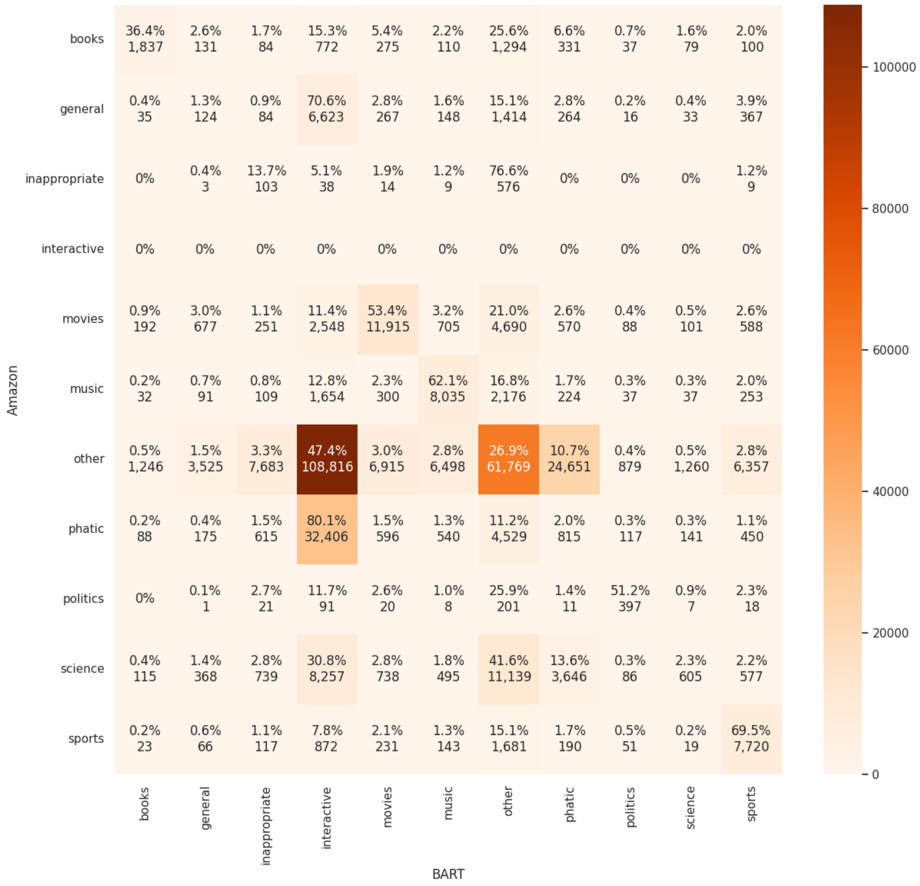
Topic level: the topics provided by the BART and CoBot topic classifiers were annotated according to the following labels.

- 2: BART topic is correct.
- 1: CoBot topic is correct.
- 0: Both BART and CoBot topics are correct.

- −1: Neither the BART nor the CoBot topics are correct.

Subtopic level: as an extension of the topic level, we annotated the same 200 random dialogues also at the subtopic level using the following labels:

- 1: The BART subtopic is correct.
- 0: The BART subtopic is not correct.



	books	general	inappropriate	interactive	movies	music	other	phatic	politics	science	sports
books	36.4% 1,837	2.6% 131	1.7% 84	15.3% 772	5.4% 275	2.2% 110	25.6% 1,294	6.6% 331	0.7% 37	1.6% 79	2.0% 100
general	0.4% 35	1.3% 124	0.9% 84	70.6% 6,623	2.8% 267	1.6% 148	15.1% 1,414	2.8% 264	0.2% 16	0.4% 33	3.9% 367
inappropriate	0%	0.4% 3	13.7% 103	5.1% 38	1.9% 14	1.2% 9	76.6% 576	0%	0%	0%	1.2% 9
interactive	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
movies	0.9% 192	3.0% 677	1.1% 251	11.4% 2,548	53.4% 11,915	3.2% 705	21.0% 4,690	2.6% 570	0.4% 88	0.5% 101	2.6% 588
music	0.2% 32	0.7% 91	0.8% 109	12.8% 1,654	2.3% 300	62.1% 8,035	16.8% 2,176	1.7% 224	0.3% 37	0.3% 37	2.0% 253
other	0.5% 1,246	1.5% 3,525	3.3% 7,683	47.4% 108,816	3.0% 6,915	2.8% 6,498	26.9% 61,769	10.7% 24,651	0.4% 879	0.5% 1,260	2.8% 6,357
phatic	0.2% 88	0.4% 175	1.5% 615	80.1% 32,406	1.5% 596	1.3% 540	11.2% 4,529	2.0% 815	0.3% 117	0.3% 141	1.1% 450
politics	0%	0.1% 1	2.7% 21	11.7% 91	2.6% 20	1.0% 8	25.9% 201	1.4% 11	51.2% 397	0.9% 7	2.3% 18
science	0.4% 115	1.4% 368	2.8% 739	30.8% 8,257	2.8% 738	1.8% 495	41.6% 11,139	13.6% 3,646	0.3% 86	2.3% 605	2.2% 577
sports	0.2% 23	0.6% 66	1.1% 117	7.8% 872	2.1% 231	1.3% 143	15.1% 1,681	1.7% 190	0.5% 51	0.2% 19	69.5% 7,720

Figure 4. Confusion matrix of the Zero-Shot BART model versus the CoBot topic classifier on the SGC5 data. The percentage and number of matches for each topic are shown.

Table 6 shows the results of the two annotation levels.

As can be seen at the topic level, the classification performed by the BART model is significantly better than the model provided by the CoBot topic classifier. In 101 cases, both models agreed (tied) in their annotation and were right, and only in 14 were they both wrong. However, in 82 cases, the topic classified by BART was better than that of the label provided by the CoBot topic classifier, whereas the CoBot topic classifier was right in 3 cases ahead of BART.

In more detail, we found that the CoBot topic classifier labeled turns as "other" topic in 71 cases. In 58 out of those 71 cases, the human label was 2; that is, the topic annotated by BART provided a better label (i.e., different from "other"); in the remaining cases, the human label was −1; that is, none of the models were correct in predicting the topic. These results mainly show that the proposed approach with a higher number of topics is able to successfully break down the turns annotated as "other" by the CoBot topic classifier, and therefore provides a better classification to the annotator.

Table 6. Results of manual annotations conducted at the topic and subtopic level on SGC5 data, with BART model versus CoBot topic classifier.

Scores	Topic Level		Subtopic Level	
	Hits	Accuracy (%)	Hits	Accuracy (%)
2	82/200	41	-	
1	3/200	1.5	143/200	71.5
0	101/200	50.5	57/200	28.5
−1	14/200	7	-	

Each time the BART model labels a turn, it returns all the possible labels it has been asked to use to classify, along with their respective associated score. The scores correspond to the probability that the label is true (confidence). During the classification of all turns, the label with the highest associated score was taken, regardless of any threshold.

Figure 5 shows the number of different topics annotated by BART and their average score associated with each topic. As can be seen, only 13 of the 18 topics were detected in the selected random subset. The missing topics are: news, photography, vehicle, weather, and other. The most represented topics were food, music, and movies. The topics with less confidence were family, finance, and education. It should be noted that “family” had a large number of occurrences; however, its average score associated was low. Regarding finance and education, they had low representation and a very low average score associated, making these the worst topics to be classified by BART.

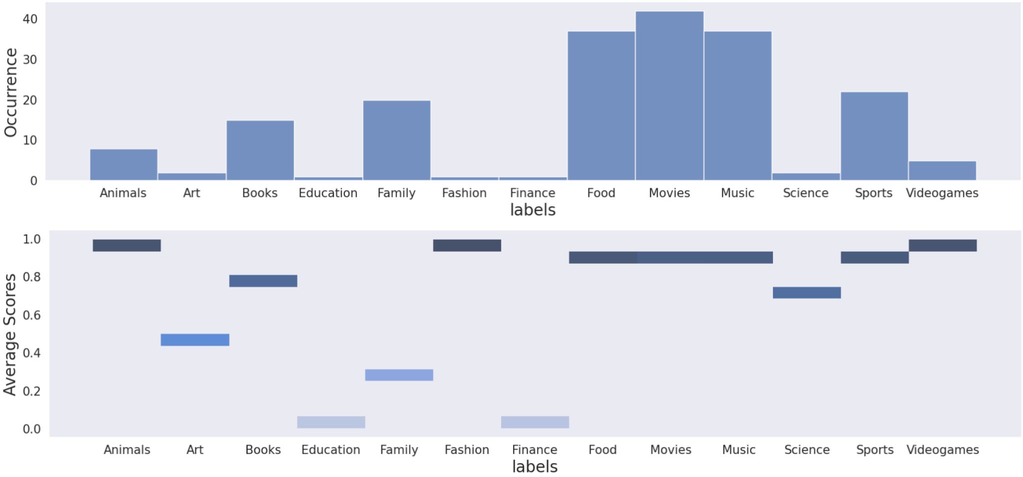


Figure 5. Topic results in the SGC5 subset of 200 dialogue turns. Upper graph: number of topics annotated by the BART classifier. Lower graph: average score associated with each topic classified by the BART model. The darker the color, the higher the average score associated with each subject, and vice versa.

Taking only the number of hits of the BART classifier at the topic level into account, we calculated how many of these hits have an associated score above a minimum threshold of 0.9 (as in Figure 3). In ∼71% of the cases it was favorable, while in ∼29% it was lower, thus demonstrating that the confidence level of the classifier is high.

At the subtopic level, the proposed BART labels were correct with an accuracy of 71.5%. This result shows that the fine-grained Zero-Shot annotations are suitable, coherent, and scalable. Of the 143 cases in which the subtopic was correct, in 142, the topic was also labeled as correct. The subtopics that were repeated the most throughout the subset were friends (family), genre (movies/music), healthy (food), and song (music). For the hit cases, only ∼31% exceeded the threshold of 0.9, showing less confidence in the choices made, although they were correct in a high percentage.

5.1.3. New Results on DailyDialog—Topic and Subtopic Level

In view of the previous results, a final test was performed using the manual annotations made in [29] on the same 1000 random turns selected from the DailyDialog dataset. For the cases where the label was "other", manual annotations were again performed, but this time taking into account the 18 proposed topics (Table 1), instead of only the original 13 topics. The results of the annotations are as follows (numbers in parentheses indicate frequency of occurrence in the dataset): food (184), finance (119), fashion (108), vehicle (56), family (39), sports (27), education (18), music (15), art (13), animals (12), weather (7), movies (6), books (4), photography (4), video games (3), news (3), science (3), and the bag-like category other (379). The generated topics were then compared with manually annotated topics.

For the subtopics, an expert performed new manual annotations. For these annotations, only the 579 turns where the BART and manual annotations matched at the topic level were taken. These turns were then annotated at the subtopic level. During annotation, only the subtopics corresponding to the annotated topic were shown to the annotator. Table 7 shows some examples of the sentences annotated in the DailyDialog subset.

Table 7. Examples of sentences and their topic and subtopic annotations. The teal color represents Human 1 and the violet color represents Human 2.

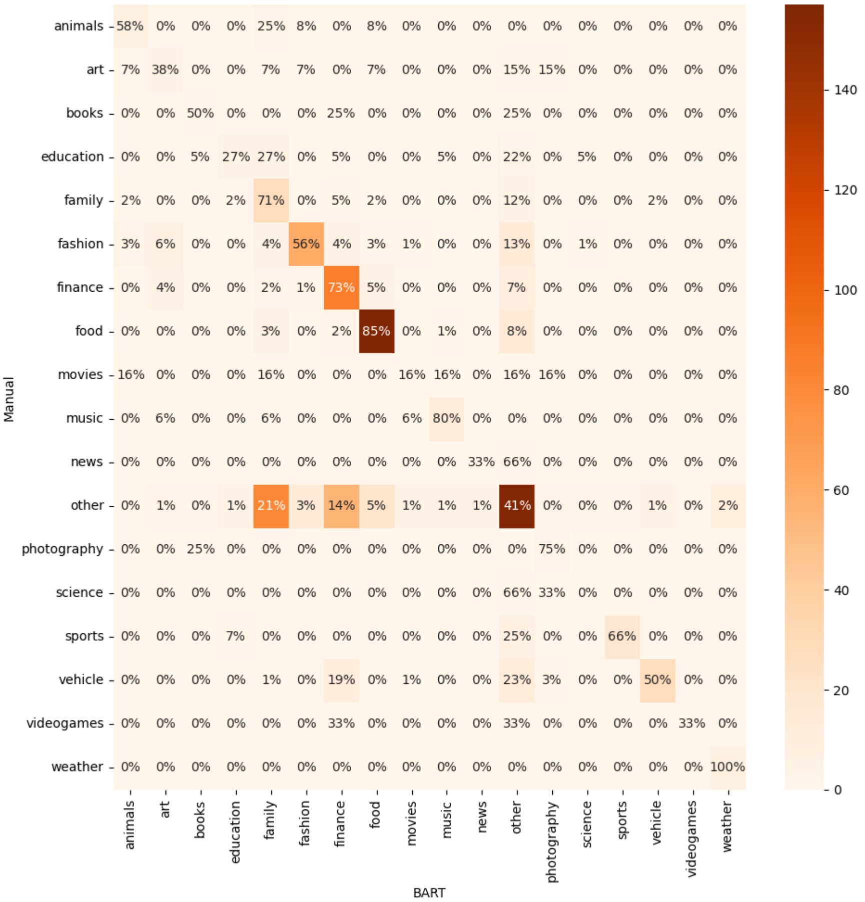
Sentence		Toipic	Subtopic
Human1: Human2:	I know. I'm starting a new diet the day after tomorrow. It's about time.	food	healthy
Human1: Human2:	I have something important to do, can you fast the speed? Sure, I'll try my best. Here we are.	vehicle	speed
Human1: Human2:	Do you know this song? Yes, I like it very much.	music	song
Human1: Human2:	Where are you going to find one? I have no idea.	other	—
Human1: Human2:	I wish to buy a diamond ring. How many carats diamond do you want?	finance	investment
Human1: Human2:	It's a kitty. Oh, Jim. I told you. No pets. It'll make a mess of this house.	animals	pets

The results in Table 8 at the topic level show an F1 score of 0.45 (57.90% accuracy), representing 579 out of 1000 turns manually classified by humans and by the BART classifier. This is a reasonable result, given that the BART model is capable of classifying up to 18 different topics (Table 1), and given the large number of different topics from which the classifier can choose, causing the error to rise. The confusion matrix is shown in Figure 6.

Table 8. DailyDialog human annotations vs. BART classification.

Model	Accuracy (%)	F1 Score	Precision	Recall
Topics	57.90	0.45	0.43	0.53
Subtopics	88.95	0.67	0.70	0.68

The results are similar to those obtained for DailyDialog in Section 5.1.1, but in this case, there are 18 topics (Table 1) versus the original 13 topics; hence, there may be a decrease in the F1 Score.



Manual \ BART	animals	art	books	education	family	fashion	finance	food	movies	music	news	other	photography	science	sports	vehicle	videogames	weather
animals	58%	0%	0%	0%	25%	8%	0%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
art	7%	38%	0%	0%	7%	7%	0%	7%	0%	0%	0%	15%	15%	0%	0%	0%	0%	0%
books	0%	0%	50%	0%	0%	0%	25%	0%	0%	0%	0%	25%	0%	0%	0%	0%	0%	0%
education	0%	0%	5%	27%	27%	0%	5%	0%	0%	5%	0%	22%	0%	5%	0%	0%	0%	0%
family	2%	0%	0%	2%	71%	0%	5%	2%	0%	0%	0%	12%	0%	0%	0%	2%	0%	0%
fashion	3%	6%	0%	0%	4%	56%	4%	3%	1%	0%	0%	13%	0%	1%	0%	0%	0%	0%
finance	0%	4%	0%	0%	2%	1%	73%	5%	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%
food	0%	0%	0%	0%	3%	0%	2%	85%	0%	1%	0%	8%	0%	0%	0%	0%	0%	0%
movies	16%	0%	0%	0%	16%	0%	0%	0%	16%	16%	0%	16%	16%	0%	0%	0%	0%	0%
music	0%	6%	0%	0%	6%	0%	0%	0%	6%	80%	0%	0%	0%	0%	0%	0%	0%	0%
news	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	66%	0%	0%	0%	0%	0%	0%
other	0%	1%	0%	1%	21%	3%	14%	5%	1%	1%	1%	41%	0%	0%	0%	1%	0%	2%
photography	0%	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	75%	0%	0%	0%	0%	0%
science	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	66%	33%	0%	0%	0%	0%	0%
sports	0%	0%	0%	7%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	66%	0%	0%	0%
vehicle	0%	0%	0%	0%	1%	0%	19%	0%	1%	0%	0%	23%	3%	0%	0%	50%	0%	0%
videogames	0%	0%	0%	0%	0%	0%	33%	0%	0%	0%	0%	33%	0%	0%	0%	0%	33%	0%
weather	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%

Figure 6. Confusion matrix of the BART Zero-Shot model versus manual annotations on the DailyDialog data subset, and for 18 different topic labels. The percentage matches for each topic are shown.

The analysis of the confusion matrix demonstrates a significant level of consensus across almost all topics. Notably, the topics of animals, books, family, fashion, finance, food, music, other, photography, sports, vehicle, and weather exhibit a higher frequency of correctly identified turns. Conversely, there are relatively fewer instances of correctly identified turns within the domains of art, education, movies, news, science, and video games. This higher percentage of successful identifications, in comparison to previously obtained results, can be attributed to the incorporation of a more refined level of granularity in the topic classification. This enhanced granularity facilitated a more comprehensive interpretation, thereby enabling the explanation of potential ambiguities or cases that would previously have been labeled as "other".

Then, at the subtopic level, the results are superior, considering only the 579 turns in which the human turn-level annotations and the BART model coincided at the topic level. In this case, in 515 cases, the manual subtopic and the one annotated by BART matched, representing a hit rate of 88.95%. In this case, the error is lower because the BART classifier has fewer options (6 subtopics on average for each topic) to classify from, unlike at the topic level, where it had to choose from 18 possibilities.

5.2. Detection of Response Inconsistencies

This section describes the results obtained in the detection of inconsistency generation in conversational agents. It also discusses the work conducted to further extend the previous results to endorse the ability of the proposed algorithm in detecting the correct number of inconsistent responses.

The results obtained for the conversational agents originally proposed are shown in Table 9. To assess the model's ability to predict the occurrence of inconsistencies, the mean square error (MSE) metric was used. The values obtained compare the average number of distinct responses manually annotated by humans and the number of clusters predicted by the proposed algorithm for each of the conversational agents. Analyzing these results, it is verified that the higher the number of parameters in their DNNs, the higher the consistency. This can be seen with the conversational models BlenderBot v2.0 400M and 2.7B. Looking at the relation between the annotated and predicted results, it can be seen that the responses obtained with the model with almost 3B parameters are more consistent compared to the 400M model. Thus, it is consistent with studies showing that a higher number of parameters in DNN models correlates with higher consistency and knowledge capability.

Analyzing the overall MSE, we obtain a 3.4. This value is very satisfactory considering the average count of distinct responses identified by the annotators in relation to the model's predicted average count of distinct responses. The Inter-Annotator Agreement (IAA), measured by Krippendorff's alpha, yielded a score of 0.74.

Table 9. Original results of the different responses generated by previous chatbots and the predicted number of inconsistencies.

Chatbot	Avg. No. Responses	Av. Predicted	MSE
BlenderBot2 (400M)	4.0 ± 1.6	4.6 ± 2.2	2.4
BlenderBot2 (2.7B)	3.7 ± 1.6	3.3 ± 1.9	2.8
DialoGPT-large	4.3 ± 2.1	3.1 ± 2.0	5.4
Seeker	4.0 ± 1.7	4.0 ± 1.7	3.1
Overall	4.0 ± 1.7	3.8 ± 2.0	3.4

The analysis performed on these results showed that high deviations were due to the generation of bland responses. In this case, despite exhibiting semantic nuances, the projections of their vector embeddings were visually located close to each other, therefore producing poorer prediction when estimating the number of clusters to the generated responses. The original study also highlights how some variations among the responses were reduced to differences in the names of the entities used within the sentence, while the remaining content of the sentence remained unchanged. For example, instances SUCH AS "I like to listen to the band The Killers." and "I like the one from the band The Who." were observed.

Many of the conclusions drawn in the original study were found to be true when looking at the results obtained with the new conversational agents.

Taking into account the reduced number of samples and aiming to increase the accuracy in the prediction values, the values shown in Tables 10 and 11 are the average values of predicting, up to 100 iterations, the number of clusters with the proposed algorithm. Table 10 shows the prediction results on the new chatbots using the same vector embedding model used in [29]. Specifically, we used the all-mpnet-base-v2 (https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (accessed on 14 July 2023)) vector embedding model. The results show that for the new models, the average number of distinctive answers given by the human annotations reduced (i.e., models are more coherent). This result is consistent with the size and quality of the new models. In our experiments, the 762M-parameter DialoGPT-large (https://github.com/microsoft/DialoGPT (accessed on 14 July 2023)) model and the 6.2B-parameter ChatGLM (https://chatglm.cn/blog (accessed on 14 July 2023)) conversational model received the highest number of variations according to the annotators. The result of ChatGLM is particularly relevant since it was originally conceived as a bilingual Chinese–English dialogue model. Many of the responses obtained with this model had to be post-processed so that they did not contain Chinese or other special characters. While maintaining lexical-semantic coherence, these particularities could be the reason for the detriment in its consistency.

Table 10. Proposed results of the different answers generated by the new chatbots and the expected number of inconsistencies with embedding encoder "all-mpnet-base-v2".

Chatbot	Avg. No. Responses	Av. Predicted	MSE
BlenderBot3 (3B)	3.7 ± 1.2	2.5 ± 0.7	2.9
ChatGLM	4.6 ± 1.4	3.8 ± 0.8	3.2
DialoGPT-large	5.1 ± 1.5	3.3 ± 0.9	5.0
GPT-4	3.4 ± 1.1	3.5 ± 0.8	1.8
Overall	4.2 ± 1.3	3.3 ± 0.8	3.3

Skimming through Table 10, the responses obtained with the GPT-4 model are those with the lowest mean square error between human annotations and predictions with an overall minimum value of 1.6. As expected, this model turned out to be the most consistent model in its responses, being the largest of all the proposed models in terms of parameters. For a better understanding of these results, the expanded questions together with the answers obtained with this model are available in Appendix A.

In view of the positive results obtained with the all-mpnet-base-v2 sentence embeddings encoder, and in order to test the robustness and dependency of the prediction algorithm when using a different vector embedding model, we performed the same approach on the same conversational chatbots and their responses, but this time using the vector representations extracted with the State-of-the-Art OpenAI text-embedding-ada-002 (https://openai.com/blog/new-and-improved-embedding-model (accessed on 14 July 2023)) encoder model. Table 11 provides the prediction results.

Table 11. Proposed results of the different answers generated by the new chatbots and the expected number of inconsistencies with embedding encoder "text-embedding-ada-002".

Chatbot	Avg. No. Responses	Av. Predicted	MSE
BlenderBot3 (3B)	3.7 ± 1.2	2.5 ± 0.8	3.5
ChatGLM	4.6 ± 1.4	3.7 ± 0.8	3.1
DialoGPT-large	5.1 ± 1.5	3.4 ± 1.0	4.7
GPT-4	3.4 ± 1.1	3.5 ± 0.7	1.6
Overall	4.2 ± 1.3	3.3 ± 0.9	3.2

A quick glance shows a slight improvement in the relationship between the annotations and the predictions on the generated responses. However, the improvements obtained with this new model are not significant. The only model whose mean square error is slightly reduced is GPT-4, by approximately 10%. As expected, this model turns out to be more consistent than the other proposed models.

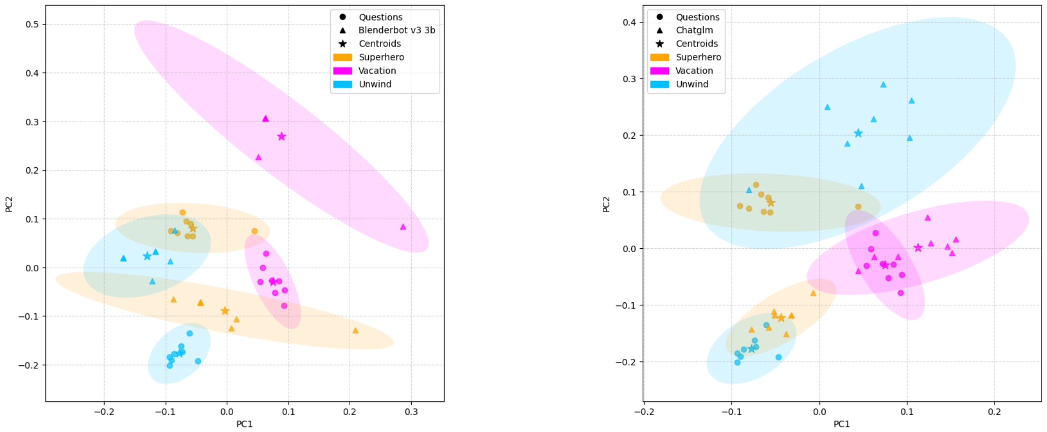
Next, for a better understanding of the results, Figure 7 shows the vector representations obtained when using the OpenAI encoder for randomly selected questions and answers from each conversational model. The figure includes error ellipses, with the same color, that group together the paraphrases and respective responses for the same type of question. This error ellipse allows us to represent an isocontour of the embedding distribution. In our case, the confidence interval displayed defines the region that would contain 99% (σ = 3.0) of all samples that could be drawn from the underlying Gaussian distribution.

Considering the small number of samples (questions and answers generated), it is essential to use confidence ellipses to better see the shape of the clusters that these vectors make up. It is interesting to note some details in these visualizations. It can be seen how for the BlenderBot3 dialogue model, specifically for the topic "Vacation", there are only three projected points. This is because this model answered with the same three sentences to the eight possible questions for this topic: "I would love to go to Hawaii.", "I would love to go to the Bahamas.", and "I would have to say Hawaii.". One can notice how, semantically, they barely differ from each other. It is also interesting to note the large size of the error ellipses for the DialoGPT-large and ChatGLM-6B models, which can visually provide confirmation of the high number of distinct answers provided by the human annotators and the proposed algorithm.

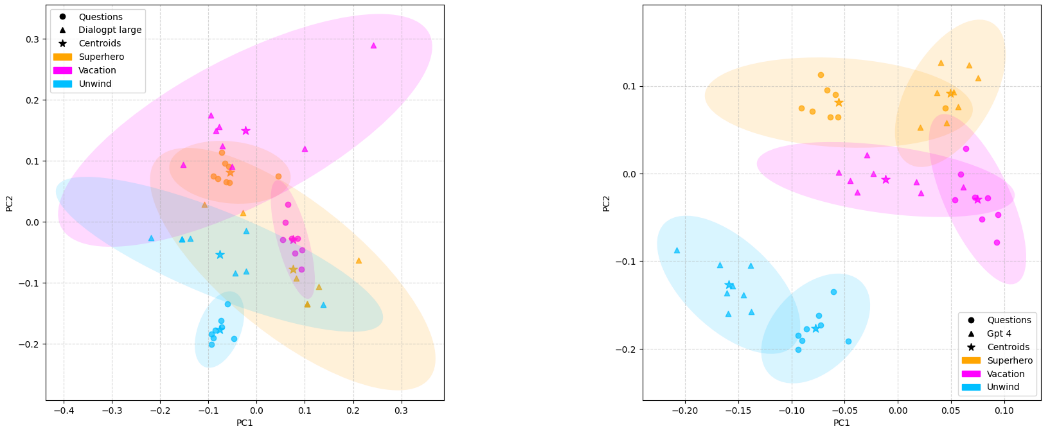
Finally, to make the human annotations and algorithm predictions clearer for the same examples shown in Figure 7, Table 12 provides the number of annotated and predicted cluster numbers for the same types of questions under discussion. Looking back at the cluster projections of the BlenderBot3 model for the ”Superhero” question, it can be checked how the human annotations and predictions correlate with what is shown in Figure 7a.

Table 12. Number of clusters manually annotated and automatically predicted using text-embedding-ada-002-based embeddings for generated chatbot’s answers.

	BlenderBot3		**ChatGLM**		**DialoGPT-large**		**GPT-4**	
Type of Question	**Human**	**Pred.**	**Human**	**Pred.**	**Human**	**Pred.**	**Human**	**Pred.**
Vacation	3.0	2.0	3.0	3.0	4.0	2.0	2.0	3.0
Superhero	2.0	2.0	5.0	5.0	5.0	4.0	5.0	4.0
Unwind	5.0	3.0	5.0	4.0	6.0	4.0	3.0	2.0



(**a**) BlenderBot3 (3B) (**b**) ChatGLM-¡ (6B)



(**c**) DialoGPT-large (**d**) GPT-4

Figure 7. Visualization of the PCA projection of embeddings and confidence ellipses of 3.0 Standard Deviations obtained with the text-embedding-ada-002 model. The graphs represent 3 groups of the new questions and answers randomly selected for each conversational model.

5.3. Latency

Table 13 shows the times required for the BART model to classify a turn in the different datasets.

Table 13. Latencies of the BART classifier calculating topics and subtopics for SGC5 and DailyDialog datasets.

Dataset	#Labels	Average Time (ms)
SGC5	11	27
DailyDialog Topic	18	39
DailyDialog Subtopic	6	19

The times shown represent the average time it takes the model to classify 20 different sentences. As expected, when the model is asked to classify the turn with a larger number of labels, the time increases since it has a greater variety of labels to choose from. However, the times for all cases are quite small, taking an average of 39 ms to choose among 18 different topics (Table 1).

Table 14 shows the times required to complete one iteration. An iteration consists of the execution of the algorithm on all the data. In our case, we have a total of 120 sentences divided into the 15 target clusters. In the same way, one iteration is the complementary execution of the k-means algorithm, obtaining the Silhouette values and the cluster prediction evaluation with the proposed inconsistency detection algorithm. The algorithm relies on the use of Silhouette values to evaluate the cluster prediction.

Table 14. K-Means and Silhouette clustering latencies for the 4 models used.

Dataset	Average Time (s)
BlenderBot3 (3B)	4.93 ± 0.05
ChatGLM	4.29 ± 0.03
DialoGPT-large	4.41 ± 0.08
GPT-4	4.40 ± 0.05

The computer used to take the measures was an Intel(R) Core(TM) i9-10900F CPU @ 2.80 GHz. The code was run on a NVIDIA GeForce RTX 3090 24 GB GPU using CUDA version 11.4.

6. Conclusions and Future Work

This research paper introduces and presents the results of a successful algorithm designed for the purpose of topic and subtopic detection in open-domain dialogues using scalable Zero-Shot approaches. Additionally, a methodology was proposed to automatically identify inconsistent responses in generative-based chatbots utilizing automatic clustering techniques. The experimental results demonstrate the efficacy of the topic detection algorithm, achieving an F1 weighted score of 0.66 when detecting 13 distinct topics and an F1 weighted score of 0.45 when detecting 18 distinct topics (Table 1). In terms of the subtopic level, a weighted F1 score of 0.67 was achieved. Moreover, the algorithm exhibits precise estimation capabilities in determining the number of diverse responses, as evidenced by an MSE of 3.4 calculated over a set of 109 handcrafted responses, 15 sets of original questions plus their paraphrases, passed to 4 small model chatbots. In the case of the 120 questions created with GPT-4, 15 question sets each consisting of 1 original and its respective 7 paraphrases, and fed into 4 State-of-the-Art chatbots, the overall resulting MSE was 3.2.

Our forthcoming research will primarily focus on two main aspects: expanding the range of high-level topics and subsequently evaluating the algorithm's performance in identifying subtopics. Additionally, we aim to include this topic and subtopic classifier in the dialogue management for the chatbot that we are using during our participation in the Alexa Socialbot Grand Challenge (SGC5). Regarding the detection of inconsistent responses, our efforts will be directed towards the development of controllable algorithms and architectures, such as TransferTransfo [36] or CTRL [50], leveraging persona profiles within these frameworks with the idea of generating more consistent responses. Furthermore, we seek to explore mechanisms to incorporate these identified inconsistencies into an automated evaluation of dialogue systems [51,52], according to the recommendations made in [20].

Author Contributions: Conceptualization, M.R.-C., M.E.-G., L.F.D., F.M. and R.C.; Methodology, M.R.-C., M.E.-G. and L.F.D.; Software, M.R.-C. and M.E.-G.; Validation, M.R.-C. and M.E.-G.; Formal analysis, M.R.-C., M.E.-G., L.F.D., F.M. and R.C.; Investigation, M.R.-C., M.E.-G., L.F.D., F.M. and R.C.; Resources, M.R.-C. and M.E.-G.; Data curation, M.R.-C.; Writing—original draft, M.R.-C.; Writing—review & editing, M.R.-C., M.E.-G., L.F.D., F.M. and R.C.; Visualization, M.R.-C. and M.E.-G.; Supervision, M.R.-C., M.E.-G., L.F.D., F.M. and R.C.; Project administration, M.R.-C. and L.F.D.; Funding acquisition, L.F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the European Commission through Project ASTOUND (101071191—HORIZON-EIC-2021-PATHFINDERCHALLENGES-01).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: This work is supported by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union". In addition, it is part of the R&D project "Cognitive Personal Assistance for Social Environments (ACOGES)", reference PID2020-113096RB-I00, funded by MCIN/AEI/10.13039/501100011033. Finally, we would like to acknowledge the support from the Amazon Alexa Prize as part of SocialBot Grand Challenge 5.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this section, we provide some examples of questions and paraphrases, as well as the different responses generated by GPT-4, followed by the number of human annotated clusters and the predicted number of distinct responses generated.

Table A1. Examples of responses and annotations for superhero topic questions. The teal color represents original sentence and the violet color represents paraphrased sentence.

Superhero Topic Questions	**GPT-4 Model Responses**
Original: Who is your favorite superhero?	My favorite superhero is definitely Spider-Man because of his relatable struggles and character development through the years.
Paraphrase: Which superhero do you hold as your favorite?	I really enjoy Spider-Man because of his relatable struggles and witty humor, plus the iconic red and blue costume.
Paraphrase: What superhero character do you admire the most?	I personally admire Spider-Man the most because of his relatability and sense of responsibility towards his community.
Paraphrase: Who is the superhero you resonate with the most?	I really resonate with Spider-Man because he's relatable, deals with everyday struggles, and always tries to do what's right.
Paraphrase: What is your preferred superhero of all time?	My all-time favorite superhero is Spider-Man because of his relatability and the way he balances his personal life with heroism.
Paraphrase: Which fictional hero holds a special place in your heart?	Harry Potter has a special place in my heart because of his bravery, growth, and the magical world he introduced me to.
Paraphrase: Who is the superhero that you find most inspiring?	My all-time favorite superhero is Spider-Man because he teaches us the importance of responsibility and resilience in tough situations.
Paraphrase: What superhero character do you consider your favorite?	I absolutely love Spider-Man because of his relatability, witty humor, and the way he deals with everyday challenges.
Average Annotation: 2.0	Average Prediction: 3.0

Table A2. Examples of responses and annotations for vacation topic questions. The teal color represents original sentence and the violet color represents paraphrased sentence.

Vacation Topic Questions	GPT-4 Model Responses
Original: What is your dream vacation destination?	My dream vacation destination would be exploring the beautiful landscapes of New Zealand and experiencing its rich culture.
Paraphrase: Where is your ultimate dream vacation spot?	My ultimate dream vacation spot would be Bora Bora, relaxing on a beautiful overwater bungalow and enjoying the crystal-clear water.
Paraphrase: What destination do you fantasize about for a perfect vacation?	I dream about relaxing on a beautiful tropical beach like Bali or Maldives, with warm turquoise water and palm trees.
Paraphrase: Which place would be your dream getaway location?	My dream getaway location would be the Amalfi Coast in Italy, with its beautiful scenery, charming towns, and delicious cuisine.
Paraphrase: If you could choose any place for a dream vacation, where would it be?	I'd absolutely love to visit New Zealand for a dream vacation, immersing myself in its breathtaking landscapes and rich culture.
Paraphrase: What is the vacation destination you yearn to visit the most?	I've always dreamed of visiting the Maldives, spending my days relaxing on the sandy beaches and swimming in crystal-clear waters.
Paraphrase: Where would you love to go on your dream vacation?	I would love to go to Bora Bora for my dream vacation, staying in an overwater bungalow and enjoying the beautiful beaches!
Paraphrase: Which location tops your list for an ideal vacation spot?	I absolutely love the idea of vacationing in Santorini, Greece, with its beautiful beaches, stunning sunsets, and unique architecture.
Average Annotation: 5.0	Average Prediction: 4.0

Table A3. Examples of responses and annotations for unwind topic questions. The teal color represents original sentence and the violet color represents paraphrased sentence.

Unwind Topic Questions	GPT-4 Model Responses
Original: How do you like to unwind after a long day?	I love to relax with a good book, watch my favorite shows on Netflix, or enjoy a soothing cup of tea.
Paraphrase: What is your preferred method of relaxation after a tiring day?	I really enjoy winding down with a good book, cozying up on the couch, and sipping on some calming chamomile tea.
Paraphrase: How do you like to de-stress and unwind following a long day?	I love curling up with a good book, watching a movie, or taking a bubble bath to de-stress after a long day.
Paraphrase: What activities do you find most rejuvenating after a busy day?	After a busy day, I find relaxing with a good book, watching a movie, or going for a walk outside to be most rejuvenating.
Paraphrase: How do you prefer to unwind and recharge after a hectic day?	I love curling up with a good book, taking a warm bath, or watching my favorite TV show to relax after a long day!
Paraphrase: What is your go-to way of unwinding after a long and exhausting day?	I love taking a warm bath, listening to soothing music, and practicing mindfulness meditation to unwind after a tough day.
Paraphrase: What helps you relax and find peace of mind after a demanding day?	I love taking a warm bath, listening to soothing music, and practicing mindfulness meditation to unwind after a tough day.
Paraphrase: How do you like to wind down and find tranquility at the end of a long day?	I like to relax by reading a good book, listening to soothing music, or taking a warm bath to find tranquility.
Average Annotation: 5.0	Average Prediction: 4.0

References

1. Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E.M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage. *arXiv* **2022**, arXiv:cs.CL/2208.03188.
2. Thoppilan, R.; Freitas, D.D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:cs.CL/2201.08239.
3. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:cs.CL/2303.08774.
4. Rodríguez-Cantelar, M.; de la Cal, D.; Estecha, M.; Gutiérrez, A.G.; Martín, D.; Milara, N.R.N.; Jiménez, R.M.; D'Haro, L.F. Genuine2: An Open Domain Chatbot Based on Generative Models. In *Alexa Prize SocialBot Grand Challenge 4 Proceedings*; 2021. Available online: https://www.amazon.science/alexa-prize/proceedings/genuine2-an-open-domain-chatbot-based-on-generative-models (accessed on 14 July 2023).

5. Hakkani-Tür, D. Alexa Prize Socialbot Grand Challenge Year IV. In *Alexa Prize SocialBot Grand Challenge 4 Proceedings*; 2021. Available online: https://www.amazon.science/alexa-prize/proceedings/alexa-prize-socialbot-grand-challenge-year-iv (accessed on 14 July 2023).
6. Hu, S.; Liu, Y.; Gottardi, A.; Hedayatnia, B.; Khatri, A.; Chadha, A.; Chen, Q.; Rajan, P.; Binici, A.; Somani, V.; et al. Further advances in Open Domain Dialog Systems in the Fourth Alexa Prize SocialBot Grand Challenge. In *Alexa Prize SocialBot Grand Challenge 4 Proceedings*; 2021. Available online: https://www.amazon.science/publications/further-advances-in-open-domain-dialog-systems-in-the-fourth-alexa-prize-socialbot-grand-challenge (accessed on 14 July 2023).
7. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical Neural Story Generation. *arXiv* **2018**, arXiv:cs.CL/1805.04833.
8. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. *arXiv* **2020**, arXiv:cs.CL/1904.09751.
9. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. *arXiv* **2020**, arXiv:cs.CL/2005.00661.
10. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning–Based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* **2021**, *54*, 3. [CrossRef]
11. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? In *Proceedings of the Chinese Computational Linguistics*; Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y., Eds.; Springer International Publishing: Cham, Swizerland, 2019; pp. 194–206.
12. Guo, Z.; Zhu, L.; Han, L. Research on Short Text Classification Based on RoBERTa-TextRCNN. In Proceedings of the 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), Kunming, China, 17–19 September 2021; pp. 845–849. [CrossRef]
13. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Association for Computational Linguistics: Vancouver, BC, Canada, 2020; pp. 8440–8451. [CrossRef]
14. Schick, T.; Schütze, H. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. *arXiv* **2021**, arXiv:cs.CL/2001.07676.
15. Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C.P.; Wang, X.Z.; Wu, Q.M.J. A Review of Generalized Zero-Shot Learning Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4051–4070. [CrossRef]
16. Tavares, D. Zero-Shot Generalization of Multimodal Dialogue Agents. In Proceedings of the 30th ACM International Conference on Multimedia; Association for Computing Machinery, MM'22, New York, NY, USA, 10–14 October 2022; pp. 6935–6939. [CrossRef]
17. Krügel, S.; Ostermaier, A.; Uhl, M. ChatGPT's inconsistent moral advice influences users' judgment. *Sci. Rep.* **2023**, *13*, 4569. [CrossRef]
18. Alkaissi, H.; McFarlane, S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **2023**, *15*, e35179. [CrossRef]
19. Dziri, N.; Milton, S.; Yu, M.; Zaiane, O.; Reddy, S. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? *arXiv* **2022**, arXiv:cs.CL/2204.07931.
20. Mehri, S.; Choi, J.; D'Haro, L.F.; Deriu, J.; Eskenazi, M.; Gasic, M.; Georgila, K.; Hakkani-Tur, D.; Li, Z.; Rieser, V.; et al. Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges. *arXiv* **2022**, arXiv:cs.CL/2203.10012.
21. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318. [CrossRef]
22. Liu, C.W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; Pineau, J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 2122–2132. [CrossRef]
23. Lowe, R.; Noseworthy, M.; Serban, I.V.; Angelard-Gontier, N.; Bengio, Y.; Pineau, J. 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)*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 1116–1126. [CrossRef]
24. Tao, C.; Mou, L.; Zhao, D.; Yan, R. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*. [CrossRef]
25. Ghazarian, S.; Wei, J.; Galstyan, A.; Peng, N. Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 82–89. [CrossRef]
26. Huang, L.; Ye, Z.; Qin, J.; Lin, L.; Liang, X. GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems. *arXiv* **2020**, arXiv:cs.CL/2010.03994.
27. Dziri, N.; Kamalloo, E.; Mathewson, K.W.; Zaiane, O. Evaluating Coherence in Dialogue Systems using Entailment. *arXiv* **2020**, arXiv:cs.CL/1904.03371.

28. Sun, W.; Shi, Z.; Gao, S.; Ren, P.; de Rijke, M.; Ren, Z. Contrastive Learning Reduces Hallucination in Conversations. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 13618–13626. [CrossRef]
29. Prats, J.M.; Estecha-Garitagoitia, M.; Rodríguez-Cantelar, M.; D'Haro, L.F. Automatic Detection of Inconsistencies in Open-Domain Chatbots. In Proceedings of the Proceeding IberSPEECH 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 116–120. [CrossRef]
30. Yin, W.; Hay, J.; Roth, D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. 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)*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3914–3923. [CrossRef]
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:cs.CL/2203.05794.
33. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *arXiv* **2019**, arXiv:cs.CL/1908.10084.
34. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:stat.ML/1802.03426.
35. Campello, R.J.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data (TKDD)* **2015**, *10*, 1–51. [CrossRef]
36. Wolf, T.; Sanh, V.; Chaumond, J.; Delangue, C. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *arXiv* **2019**, arXiv:cs.CL/1901.08149.
37. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 936–945. [CrossRef]
38. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; pp. 986–995.
39. Khatri, C.; Hedayatnia, B.; Venkatesh, A.; Nunn, J.; Pan, Y.; Liu, Q.; Song, H.; Gottardi, A.; Kwatra, S.; Pancholi, S.; et al. Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize. *arXiv* **2018**, arXiv:cs.CL/1812.10757.
40. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-Enhanced Bert with Disentangled Attention. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
41. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:cs.CL/1910.13461.
42. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2020**, arXiv:cs.CL/1910.01108.
43. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv* **2020**, arXiv:cs.CL/1911.00536.
44. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E.M.; et al. Recipes for Building an Open-Domain Chatbot. *arXiv* **2020**, arXiv:cs.CL/2004.13637.
45. Xu, J.; Szlam, A.; Weston, J. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. *arXiv* **2021**, arXiv:cs.CL/2107.07567.
46. Komeili, M.; Shuster, K.; Weston, J. Internet-Augmented Dialogue Generation. *arXiv* **2021**, arXiv:cs.AI/2107.07566.
47. Shuster, K.; Komeili, M.; Adolphs, L.; Roller, S.; Szlam, A.; Weston, J. Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion. *arXiv* **2022**, arXiv:cs.CL/2203.13224.
48. Zeng, H. Measuring Massive Multitask Chinese Understanding. *arXiv* **2023**, arXiv:cs.CL/2304.12986.
49. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. *arXiv* **2022**, arXiv:cs.CL/2103.10360.
50. Keskar, N.S.; McCann, B.; Varshney, L.R.; Xiong, C.; Socher, R. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv* **2019**, arXiv:cs.CL/1909.05858.
51. Zhang, C.; Sedoc, J.; D'Haro, L.F.; Banchs, R.; Rudnicky, A. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. *arXiv* **2021**, arXiv:cs.CL/2111.02110.
52. Zhang, C.; D'Haro, L.F.; Friedrichs, T.; Li, H. MDD-Eval: Self-Training on Augmented Data for Multi-Domain Dialogue Evaluation. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 11657–11666. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.