future internet

MDPI

Article

Enhancements in BlenderBot 3: Expanding Beyond a Singular Model Governance and Boosting Generational Performance

Ondrej Kobza †, David Herel †, Jan Cuhel †, Tommaso Gargiani †, Jan Pichl †, Petr Marek †, Jakub Konrad † and Jan Sedivy *

Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University, 156 00 Prague, Czech Republic; kobzaond@fel.cvut.cz (O.K.); hereldav@fel.cvut.cz (D.H.); cuheljan@fel.cvut.cz (J.C.); pichljan@fel.cvut.cz (J.P.); marekp17@fel.cvut.cz (P.M.); konrajak@fel.cvut.cz (J.K.)

* Correspondence: jan.sedivy@cvut.cz

† These authors contributed equally to this work.

Abstract: This paper provides a pioneering examination and enhancement of generative chat models, with a specific focus on the BlenderBot 3 model. Through meticulous interaction with a diverse set of human participants, we dissected the fundamental components of these models, unveiling several deficiencies, including long-term memory and entity recognition. Leveraging these insights, we engineered refined, streamlined iterations, culminating in a chatbot that transcends the capabilities of all existing models. Our work follows Occam's razor principle and proves that, for tasks with relatively low complexity, using large overparameterized models instead of smaller ones does not bring significant benefits but increases latency, which may result in a lowered overall user experience. In upholding our commitment to transparency and the progression of shared knowledge, we have made our improved model universally accessible through open-source distribution.

Keywords: BlenderBot; ChatBot; DeBERTa; LLM; Transformer

check for updates

Citation: Kobza, O.; Herel, D.; Cuhel, J.; Gargiani, T.; Pichl, J.; Marek, P.; Konrad, J.; Sedivy, J. Enhancements in BlenderBot 3: Expanding Beyond a Singular Model Governance and Boosting Generational Performance. *Future Internet* **2023**, *15*, 384. https://doi.org/10.3390/fi15120384

Academic Editor: Michael Sheng

Received: 5 October 2023
Revised: 13 November 2023
Accepted: 25 November 2023
Published: 28 November 2023

1. Introduction

Recent years have seen a remarkable acceleration in the advancement of language models, pushing the boundaries of the state-of-the-art (Brown et al. [1], Ouyang et al. [2], OpenAI [3]). This progress has significantly impacted dialogue applications, where the interaction between humans and these models plays a pivotal role. These models, such as BlenderBot 3 (BB3) (Shuster et al. [4]), have made significant strides in improving the quality of interaction, exhibiting an impressive ability to understand and respond in a human-like manner. The current state-of-the-art in the field of open-domain chatbots involves using a large language model (LLM) to perform various steps before the generation of a final utterance—typical steps are to decide whether to use external knowledge, knowledge extraction, and long-term memory inclusion (Shuster et al. [4], University of California [5], Stevens Institute of Technology [6]). However, like any technology, there is always room for further enhancement and innovation.

In this work, we extensively analyze the fundamental elements of the BlenderBot 3 chatbot, evaluating both their conversational capabilities, including engagement and knowledge, and their adherence to behavioral standards. Notably, we are focusing on the BlenderBot 3 version with the three-billion-parameters (3B) encoder–decoder generative model, as it is easily accessible to the public due to its low computational requirements with respect to its larger versions. However, our work also applies to other LLMs, such as Llama (Touvron et al. [7]).

In our commitment to transparency and the advancement of collective knowledge, we have made the conscious decision to open-source our work (https://github.com/kobzaond/Enhancements_in_BlenderBot_3, accessed on 2 October 2023). We believe that by making our improved model accessible to everyone, we can foster a collaborative environment that propels further improvements in this field.

Our paper begins with an examination of the relevant literature and research, following which we delve into our comprehensive study of the BlenderBot 3 pipeline. The ensuing two sections shed light on our methodological approach aimed at overcoming various inconsistencies detected during the study. The concluding segment furnishes the outcomes of our proposals, thereby substantiating the assertions we made earlier in the Analysis section.

2. Related Work

A number of studies have been conducted in the area of open-domain chatbots. Roller et al. [8] introduced a process for developing these chatbots, leading to the creation of BlenderBot's first version, the Blended Skill dataset, and a rudimentary method for the integration of external knowledge from Wikipedia dumps. Contrary to past practices focusing on enhancing performance via parameter amplification, BlenderBot 1.0 improves conversational capabilities via the utilization of blended task datasets, including those related to empathy, persona, and knowledge.

The second iteration, BlenderBot 2 (https://parl.ai/projects/blenderbot2/, accessed on 16 July 2021), adopts the knowledge-based generation concepts suggested by Komeili et al. [9], such as Fusion-in-Decoder (FiD) (Izacard and Grave [10]), Retrieval-Augmented-Generation (RAG) (Lewis et al. [11]), and Dense Passage Retrieval (DPR) (Karpukhin et al. [12]). This version is built upon their unique dataset, Wizard of Internet (WoI), and incorporates the Multi-Session Chat dataset (MSC, Xu et al. [13]), yielding a tool for storing user preferences in long-term memory for multi-session dialogues. The WoI dataset is inspired by the Wizard of Wikipedia dataset (Dinan et al. [14]) but captures more topics and therefore provides better generalization. According to Shuster et al. [15], incorporation of external knowledge (or memory) will ground the generation process and reduce hallucinations (making up the model's own 'facts'). Hallucination of LLMs is a well-known problem elaborated in many studies (e.g., Ji et al. [16], Shuster et al. [15], or Zhang et al. [17]).

Furthermore, Lee et al. [18] highlighted multiple challenges in BlenderBot 2, ranging from dataset issues to long-term memory problems, complications due to external knowledge integration, internet search capability, and inappropriate responses. They suggested a number of improvements, many of which were implemented in BlenderBot 3 Shuster et al. [4].

The third version improved knowledge integration, using a method proposed by Shuster et al. [19]. It introduced several classification tasks, notably determining when to conduct a search or utilize long-term memory and the addition of an entity extraction component. Notably, all of these proposed tasks were managed by one single generative model. BlenderBot 3 comes in pre-trained transformer models (Vaswani et al. [20]), with varieties ranging from a 3B encoder–decoder version (same architecture as proposed by Lewis et al. [21]) to a 175B decoder-only model. Due to computational demands, our research focuses predominantly on the 3B model. We provide a detailed description of BlenderBot 3 in Appendix A.

Recently, novel LLMs were released, boosting the overall NLP field. Touvron et al. [7] proposed the Llama model and Penedo et al. [22] proposed Falcon LLM. These models can be utilized in the open-domain chatbots either as they are or integrated within a pipeline. Stanford University [23] utilized Falcon and BlenderBot 3 in their open-chat system, using Falcon 40B for pregenerating dialogue trees and BlenderBot's model to generate utterances in real-time (addressing the major drawback of Falcon—i.e., latency). University of California [5] and Stevens Institute of Technology [6] incorporated Llama-based models into their chatbots, achieving top results in the Amazon SocialBot Grand Challenge competition (https://www.amazon.science/alexa-prize/socialbot-grand-challenge, accessed on 16 August 2021).

Following the success of instruction-based models, Taori et al. [24] introduced the Alpaca LLM, and Chung et al. [25] introduced several instruction-based models building upon T5 (Raffel et al. [26]) and PaLM (Chowdhery et al. [27]).

In the realm of conversational settings, there has been a significant amount of work dedicated to creating innovative conversational datasets. For instance, Rashkin et al. [28] presented the Empathetic Dialogues, while Radlinski et al. [29] introduced CCPE. Additionally, Henderson et al. [30] established a repository of conversational datasets.

Numerous studies demonstrate the ability of comparatively smaller models to perform competitively against larger Language Learning Models (LLMs) on the General Language Understanding Evaluation (GLUE) tasks (Wang et al. [31], Jiao et al. [32], Clark et al. [33]). This implies that deploying smaller models may be more suitable for certain challenges as opposed to utilizing excessively large models.

3. Analysis

Our examination focused on specific modules of the BB3 chatbot, where we aimed to tackle significant performance deficiencies along with latency and computational complexity problems. Note that the major generative model in our refined pipeline can be with Llama, Falcon, or any other LLM. However, in this work, we focus on the BB3 3B model due to its relatively low memory and computational requirements and, therefore, use the original BB3 model as a reference for our comparisons.

Twelve linguist participants, selected through interviews conducted by experts from the Alquist group (http://alquistai.com/, accessed on 14 October 2016) and PromethistAI (https://promethist.ai/, accessed on 3 October 2022), were assigned the task of evaluating the BB3 chatbot under a predefined setting. As illustrated in Tables 1 and 2, the resulting evaluations should be interpreted in comparison to the baseline analysis of the reference 3B BB3. During the interviews, the participants analyzed several conversations between the Alquist chatbot and Alquist team members. Based on this analysis, the experts then selected the twelve most promising linguists.

Table 1. Average rating over 100 conversations with the original BB3 model versus the modified BB3 (after fine-tuning). The quality of the conversation was ranked on a scale (1–5). Depending on the rating category, an upward arrow (↑) indicates that higher is better; a downward arrow (↓) indicates that lower is better.

Category	Setting				
	BB3 Original	**Our BB3 Version**	**BB3 No Memory**	**BB3 No Entity**	**Vanilla BB3**
Overall Feel ↑	2.4	2.61	2.38	2.44	2.36
Repetition ↓	2.67	2.12	2.31	2.42	2.25
Hallucination ↓	2.13	2.06	2.15	2.11	2.23
Engagingness ↑	2.87	2.82	2.53	2.47	2.39

Table 2. Frequencies of particular flaws given a particular setting of the chatbot. Note that the irrelevancy (including nonsense) flaw subsumes contradictions.

Category	Setting				
	BB3 Original	**Our BB3 Version**	**BB3 No Memory**	**BB3 No Entity**	**Vanilla BB3**
Repetitions	24.11%	15.95%	21.02%	23.77%	19.33%
Irrelevancy	11.56%	10.52%	11.82%	11.69%	11.78%
Contradictions	6.08%	6.14%	6.15%	6.17%	6.68%
Hallucinations	2.21%	2.06%	2.24%	2.19%	3.12%

The outcomes of the analyses in this section are mostly preemptive findings, which are then further elaborated in the Method and Results Sections 4 and 5. We first elaborate on the flaws of the whole BlenderBot 3 system, and afterward, we dive into the analysis of particular modules of the BB3 pipeline.

3.1. General Flaws

During the testing of BB3, we identified the following problems (for each of the problems, we provide an example in Appendix B):

1. Sometimes, it does not foster the conversation more deeply, resulting in shallow conversations. This resembles more superficial discussions, contrasting with the natural human tendency to delve deeper into particular issues. Rapidly changing subjects can sound unnatural and potentially irritating.
2. Repetitions: i.e., the chatbot sometimes replies with semantically the same utterances as it did in the conversation before. This significantly reduces the overall feel of conversations with the chatbot.
3. Occasionally, the chatbot produces irrelevant outputs, disrupting the conversation's consistency.
4. The chatbot's responses are frequently too terse. While a chatbot should indeed be engaging, a single-word response is more likely to hinder the conversation rather than facilitate it.
5. Hallucinations, i.e., making up the chatbot's own 'facts,' which are not true. The standard solution is to condition the generation by a 'knowledge' so that the generative model uses the knowledge in its input context and, therefore, does not have to make up its own 'facts' [15]. Since BlenderBot 3 specifically adopts this approach, the frequency of hallucination is relatively low; however, the flaw is not eliminated entirely.
6. We also found instances of contradictions and failure to comprehend user input. The prevalence of this issue often correlates with the model's size, so we will not delve into this problem in depth here.
7. Lastly, we found the chatbot's high latency an issue, as waiting several seconds for a response can be inconvenient.

3.2. Relevant Entity Recognition

BlenderBot 3 extracts a pertinent entity from the comprehensive input context in instances where neither memory nor knowledge is necessitated. This extraction of the relevant entity underpins the ultimate response, potentially assisting the generative model to pinpoint the 'centroid' of the conversation, thereby improving the quality and relevance of the generated responses.

Nonetheless, our initial manual testing revealed that integrating the entity into the model's context sometimes results in semantic repetitions. In other words, the model may produce an utterance that is semantically identical to one previously present in the context history or cycle back to previously discarded topics. Another clear limitation of this method is the increase in computational complexity, which adversely affects the overall latency.

3.3. Long Term Memory

Recently, research has found that providing long-term memory functionality to generate and store memories extracted dynamically from conversations is effective in improving the conversation quality of chatbots (Xu et al. [13]). However, the current state-of-the-art architectures for memory incorporation, as seen in chatbots like BB3, are not flawless. Atkins et al. [34] examined the possibilities of misinformation injection through long-term memory, finding out that when a chatbot is questioned on the misinformation topic, it increases the magnitude of misinformation generation by more than two times.

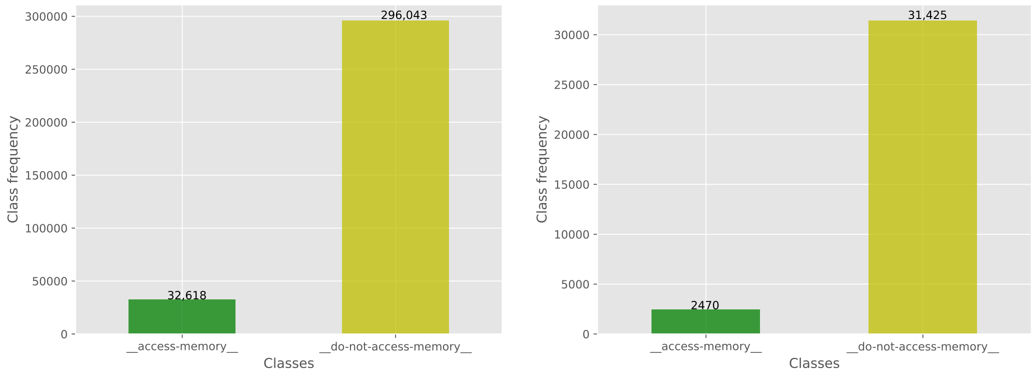
Furthermore, our findings suggest that employing memory in the BB3 setup tends to increase the percentage of repetitive outputs from the generative model, as shown in Table 2. Upon further empirical investigation, we learned that incorporating memory into the input tends to enhance output utterance quality, mainly when a user refers to previous turns. However, one negative side effect is the model's tendency to dwell on or revert back to a specific topic despite a user's desire to shift the conversation elsewhere.

3.4. Memory and Search Decision

The search and memory decision classifications are crucial in obtaining relevant information about the current topic and getting a past context. In these steps, the modules are tasked with deciding whether it is necessary to conduct an internet search or memory retrieval to get the main generative model relevant pieces of information concerning the current conversation.

It is important to note that while regarding memory decision, we are interested in false positives and false negatives equally, concerning the search decision module, we must be cautious about false positives (predicting that the search is required even though it is not). This is because the internet search is a rather time-consuming operation.

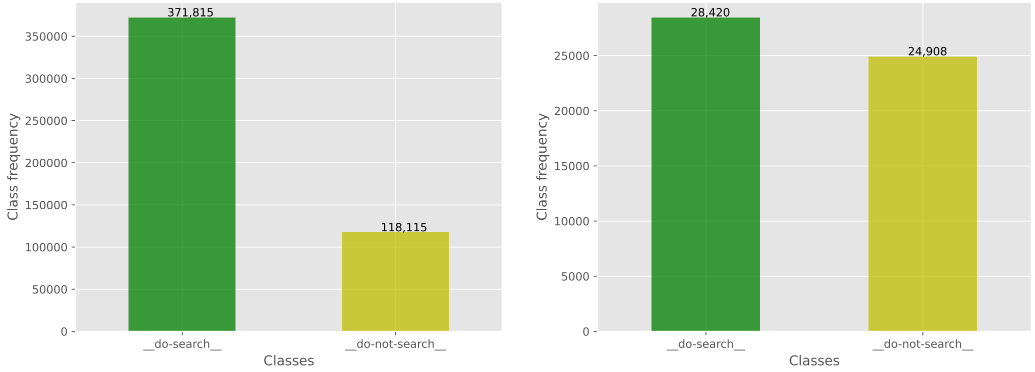
Regarding memory decision, the training (and validation) datasets utilized for BB3 on this task, as depicted in Figure 1a,b, are significantly imbalanced. This inherent imbalance inevitably influences the model's efficiency on this task, reducing the overall performance due to inaccurate memory decision classifications and resulting in the aforementioned flaws.



(**a**) Label frequency of train dataset (**b**) Label frequency of validation dataset

Figure 1. Label frequencies: train and validation datasets for memory decision task.

Similarly to the memory decision training dataset, the classes for the search decision training data are imbalanced, however in the validation set, they are not (as shown in Figure 2a,b). This imbalance in the training set potentially leads to an increased risk of a substantial bias during the training phase. At the same time, the evaluation dataset should reveal such bias during the training process.



(**a**) Label frequency of train dataset (**b**) Label frequency of validation dataset

Figure 2. Label frequencies: train and validation datasets for search decision task.

3.5. Query Generation

In the query generation task, the model is tasked with generating a query for an internet search based on the given user utterance. The relevance and specificity of the generated query are of the highest importance. Irrelevant queries may yield unrelated information from the search engine, leading the generative model to provide a non-relevant response to the user's utterance, whereas queries not specific enough will result in retrieved knowledge that is too general and prevent the chatbot from diving more deeply into the currently discussed topic.

3.6. Knowledge Extraction

BlenderBot 3 is remarkably proficient in amalgamating external knowledge with the given context before producing the final response. Initially, it formulates a query that is submitted to a search engine such as Mojeek (https://www.mojeek.com/, accessed on 23 November 2013). This then retrieves k documents, which are processed by the BB3 model using FiD. Based on our empirical examinations, this method is reasonably practical. However, it does have a notable drawback: latency. The entire sequence of knowledge extraction can require over a second, significantly slowing down the response time.

3.7. Final Response Generation

The final response generation of BB3 is marked by a series of challenges. A primary issue, as evidenced in Table 1, is the system's tendency towards repetition. It appears that BB3 often reformulates previously generated responses, thereby recycling them in its conversation flow.

Moreover, we empirically found instances of contradictions and hallucinations within the system's outputs, indicating a lack of consistency and accuracy. Despite attempts to rectify these errors through the integration of additional models, the issues persist. A detailed overview of these findings, complete with examples, is provided in Appendix B.

We posit that these challenges might stem from the language model's limited exposure to conversational data. To address the issue of repetition, we have implemented a strategy involving negative sampling. The specifics of this approach are elaborated on in Section 4.7.

4. Method

In our proposed improvements, we address the aforementioned bulk of drawbacks. This includes offering novel datasets, fresh models, and a reimplementation of the BB3 pipeline aimed at improving the identified deficiencies.

Our guiding philosophy is the Occam's Razor principle. We simplified the existing pipeline using smaller, more efficient models that have superior performance. This refined pipeline has been designed to simplify operations, boost performance, and curtail latency substantially.

We deployed our modified version of BB3 on the Amazon AWS platform using Nvidia Tesla T4 GPUs (https://www.nvidia.com/en-us/data-center/tesla-t4/, accessed on 12 September 2018), which were used as a reference point for latency measurements.

To address the imbalances in BlenderBots' evaluation datasets, which primarily come from the same sources as their training equivalents (and are likely derived from the same or very similar probability distribution, potentially introducing bias), we created our own test datasets. These were explicitly targeted at decision making in search, memory recall, and query generation areas. Through these new datasets, we aim to showcase the true performance potential of the respective models.

4.1. Test Datasets

The creation of each test dataset was carried out following a three-stage process. In the first step, we manually scrutinized the inputs and corresponding outputs from all modules during the BlenderBot 3 (empirical) testing phase, aided by inputs from our twelve Turkers.

Through these investigations, we detected specific problematic patterns. The associated user utterances were then utilized to construct the initial seed.

The second stage consisted of manually expanding the initial seed; our goal was to encompass a wider range of instances rather than solely focusing on the several identified challenging patterns.

The final phase hinged on the advanced abilities of OpenAI's ChatGPT (https://chat.openai.com, accessed on 30 November 2022) to further bolster the dataset refined in Stage 2. We supplied the model with ten manually picked representative samples from the preceding stages and asked ChatGPT to produce n additional samples. Each test set consists of a total of 300 samples.

4.2. Quality Measures

We evaluated the overall effectiveness of BB3 and the cumulative influence of our proposals through two distinct methodologies:

1. User Ranking: Our user group, consisting of twelve Turkers, was tasked with interacting with the bot under varying conditions and rating each bot's response on a scale of 1 to 5. Four attributes were identified for rating. For the criteria *overall feel* and *engagingness*, a higher score is deemed better. For the criteria of *repetition* and *hallucination*, a lower score is favorable. Each attribute was rated independently, meaning just one attribute was evaluated in a single session. We collected over 2000 turns in total.
2. Frequency of Flaws: The Turkers were provided with the conversations that were obtained during the previous testing phase and were requested to assign labels to each utterance made by the bots. The assessment involved identifying five specific flaws in the bot's utterances: *repetitions, irrelevant and nonsensical outputs, contradictions, and hallucinations*. The labeling task was binary, meaning that for each flaw mentioned (e.g., repetition), the Turkers were required to label each utterance made by the bots as either containing the flaw or not containing it.

It is worth noting that the Turkers were directed to concentrate on a single attribute or flaw per session for both methodologies. We opted for this setup to ensure the Turkers' full attention to every single attribute and flaw.

4.3. Relevant Entity Recognition

Following the empirical scrutiny delineated in the previous section, we proposed the possibility of excluding this process entirely. We inferred that this action may not degrade the overall output quality. On the contrary, it could streamline the framework, minimize the occurrence of some defects, enhance latency, and decrease computational complexity. According to the outcomes (Tables 1 and 2), there are not any statistically significant positive effects on performance, and therefore, in terms of latency improvement, we decided to omit this module in our implementation.

4.4. Memory and Search Decision

Our assumption was that the complexity of these two classification tasks could be sufficiently addressed using a comparatively smaller Transformer encoder. To this end, we sought to fine-tune smaller models such as BERT-tiny and ELECTRA-small and thus follow Occam's Razor principle. Our efforts yielded several models where our DeBERTa-xsmall shows superior performance with respect to all other models, including BB3 3B. Based on our results, which are presented below, we conclude that using a large overparameterized model for this particular problem does not bring any benefit.

4.5. Knowledge Extraction

In contrast to utilizing a search engine such as Mojeek, as is the case with BB3, we suggest the employment of a robust QA engine like Amazon EVI (i.e., Alexa's semantic knowledge graph, which is also used in production to answer Alexa questions). The advantage is that it retrieves key information directly, eliminating the need for a list of documents and thus obviating the requirement for a knowledge extraction module. This modification substantially improves latency. The underlying reasons are two-fold: firstly, EVI surpasses Mojeek in speed by an average of 380 ms, and secondly, the omission of the knowledge extraction module accelerates the entire system by over 500 ms. Consequently, the total system speed is enhanced by approximately 1 s.

However, acknowledging that EVI does not consistently return an answer, we also recommend integrating a semantic similarity retrieval from a Wikipedia dump, pipelined to FiD, to serve as a backup solution. We call this overall knowledge retrieval system *APIHUB*.

4.6. Query Generation

Based on our empirical findings, the BB3 model usually generates high-quality search queries. However, considering its relatively high latency, we sought to develop an alternative model with competitive query generation performance but a significantly reduced latency. Therefore, we designed query generators based on the FLAN-T5 base and large models, both of which have fewer parameters than the BB3 model (FLAN-T5-Base has 250M parameters, FLAN-T5-Large has 780M parameters, while BB3 has 3B parameters). Note that latency is critical in this step, as the workflow *query generation -> search -> knowledge extraction* cannot be parallelized, i.e., the 'knowledge extraction' workflow is the main bottleneck in terms of overall latency.

4.7. Improving Final Response Generations

Our comprehensive analysis of the general BB3 3B generative model itself led us to two significant findings. Firstly, the language model's limited exposure to conversational data resulted in less engaging responses. Secondly, the system exhibited a propensity for repetitive and contradictory responses.

In response to these challenges, we devised a two-pronged strategy. Initially, we fine-tuned the original model using a more extensive conversational dataset to enhance the quality of responses. These datasets were CCPE, Empathetic Dialogues, and other conversational data from Reddit and OpenSubtitles.

To address the issue of contradictions in the language model's outputs, we employed negative sampling. The crux of this method is the creation of a dataset comprising examples of conversations riddled with contradictions, hallucinations, and repetitions. This dataset serves as a reservoir of negative examples for the model. Subsequently, we modified the loss function to minimize inverse cross entropy (Equation (1)), thereby encouraging the model to unlearn these problematic patterns.

$$loss = \frac{1}{-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})} \tag{1}$$

4.8. Parallelization

The original BlenderBot 3 from the ParlAI library lacks an inference pipeline supporting parallelization. Our proposed enhancement enables the concurrent execution of query generation, search, and memory decision tasks. The obtained results are processed in parallel within the knowledge and (in certain settings) memory retrieval. Knowledge querying utilizes Amazon EVI or a semantic Wikipedia search combined with FiD. Memory retrieval via semantic similarity occurs alongside extraction by FiD. User input, context, knowledge, and memory are integrated into the 3B generative model, yielding the final response.

Our new, compact models ensure manageable deployment with the proposed paralleled pipeline since their small size ensures low memory and computational requirements. Conversely, implementing the 3B model for every task would make parallelization more memory and computation-intensive.

5. Results

In this section, we present the results of our experiments regarding specific modules and the overall performance of the chatbot across different settings/modules. Furthermore, we analyze the effects of various settings and the impact of our proposed solutions on the major flaws identified during the analysis of BB3.

Based on our comprehensive analysis, we have significantly improved the performance of this chatbot. Ineffective models, as identified by our analysis, have been eliminated. We have successfully increased the chatbot's speed and efficiency through various enhancements, such as the integration of APIHUB (described in Section 4.5) and the simplification of certain models. Our modifications were motivated by the famous Occam's razor principle—i.e., we simplified the chatbot and obtained better performance—our modified chatbot outperforms the original BlenderBot 3 by almost 9% in terms of overall quality and is more than three times faster.

We evaluated the performance concerning search and memory decision activities through standard metrics such as accuracy and F1 score. Concerning query generation, we provide SacreBLEU scores along with two manually acquired scores: Accuracy and Weak Accuracy, as explained in Table 3, which displays the comparison of the query generation task, where we compare our two best checkpoints with the original BlenderBot 3 model. In this case, we were unable to reach a definitive conclusion on whether our new models are generally better or worse than BB3. However, considering that latency is a significant bottleneck for the chatbot, we decided to implement the FLAN-T5-base-based model in our revised 'BlenderBot' system.

Table 3. We evaluated the performance of the BB3 model and our fine-tuned models (FLAN-T5 and FLAN-Alpaca) in query generation using a test dataset. We assessed the results based on three criteria: exact match, strict semantic match (in which the generated query is not an exact match but holds the same semantic meaning as the label), and relevance (where the generated query pertains to the topic but may not be as precise). It is important to note that a query with an exact match or strict semantic match is also considered relevant. The accuracy metric is a combination of exact match and strict semantic match, while weak accuracy corresponds to relevance.

Model	Model Size	Accuracy	Weak Accuracy	SacreBLEU (EVAL)	Speed Up
BB3	3B	73.6	96.2	12.1	1x
FLAN-Alpaca-large	780M	56.6	98.1	13.6	2.2x
FLAN-T5-base	250M	54.7	98.1	12.2	4.3x

The performance of our classification models is shown in Table 4. The evaluation datasets, referred to as eval datasets, are obtained by merging the evaluation datasets from all the data used for each specific task by BB3. On the other hand, the test datasets are our newly designed datasets (described in Section 4.1). Figures 3–6 show confusion matrices of BB3 and the fine-tuned classification models for Memory decision and Search decision.

The impact of specific pipeline settings (including our enhancements) on major flaws such as repetitions and contradictions is presented in Table 2. The results were obtained through the manual labeling of the chatbot's output utterances. We consider several settings to test the impact of long-term memory, dominant entity, external knowledge, as well as our our new proposed setting, which has superior performance in three out of the four categories. Note that 'Vanilla BB3' refers to a setting where all side modules (external knowledge, memory, and entity) are discarded, and the 3B model generates a response solely based on the user's utterance and context.

Table 1 provides an overview of the overall performance of the chatbot under various settings, as evaluated by Turkers on a scale from 1 to 5. The reported scores include Overall Feel and Engagingness (higher score is better), as well as Repetition and Hallucination (lower score is better). We hypothesize that for some categories, especially the Overall Feel, latency could make a certain impact on the resulting scores. The results presented in Table 1 clearly show a significant improvement in our implementation in the 'Overall Feel' and Repetition categories while showing similar performance to the original BB3 in the 'Hallucination' and 'Engagingness' categories. It is important to note that although the Turkers were instructed to be as objective as possible, the results may slightly deviate from the hypothetical rankings made by a different group of Turkers; Tables 5 and 6 provide insights into the differences of rankings among Turkers.

Table 7 shows an overview of latencies of particular chatbot's pipelines and relative speed with respect to the original BB3 chatbot. Our enhanced chatbot has an estimated latency of 1.5 s and is more than three times faster than BB3.

Table 4. Performance of BB3 model and our fine-tuned models on two classification tasks: search and memory decision.

Model		**Accuracy [%]**		**F1 [%]**		**Speedup**
Task Name	**Model Name**	**Eval**	**Test**	**Eval**	**Test**	**Test + Eval**
search decision	BB3	84.0	63.0	86.2	71.3	1x
search decision	ELECTRA-small	88.1	69.7	89.3	73.5	50x
search decision	BERT-tiny	90.3	78.0	90.6	76.4	192x
search decision	DeBERTa-xsmall	92.2	85.0	92.3	82.6	17x
memory decision	BB3	76.1	69.7	36.9	81.6	1x
memory decision	ELECTRA-small	87.3	59.8	42.5	59.8	50x
memory decision	DeBERTa-xsmall	92.1	72.0	92.1	82.7	17x

For enhanced lucidity, we also include the standard deviations of ratings from the Turkers (Tables 5 and 6). This not only illustrates the degree of consensus or disparity amongst the Turkers but also serves as an indicator of the impartiality or subjectivity embodied in the evaluation.

Table 5. Standard deviations between Turker's rating with respect to Table 1. Depending on the rating category, an upward arrow (↑) indicates that higher score is better; a downward arrow (↓) indicates that lower is better.

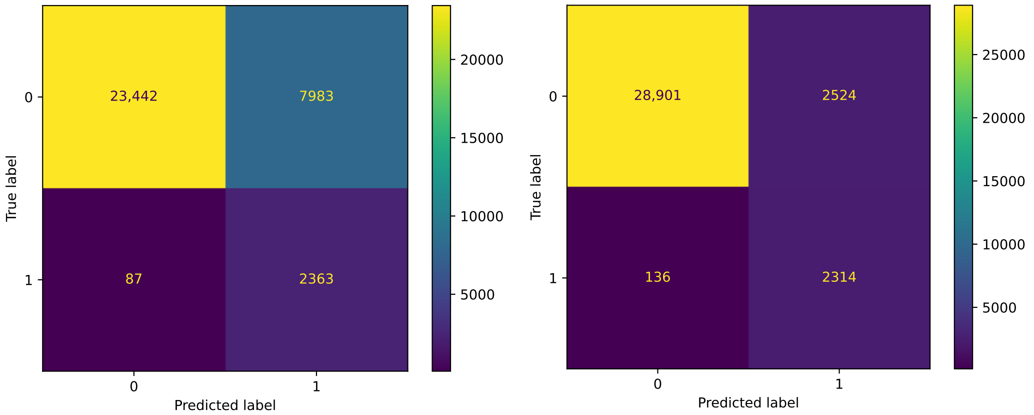
Category	**Setting**				
	BB3 Original	**Our BB3 Version**	**BB3 No Memory**	**BB3 No Entity**	**Vanilla BB3**
Overall Feel ↑	0.54	0.36	0.39	0.45	0.47
Repetition ↓	0.45	0.51	0.38	0.31	0.20
Hallucination ↓	0.65	0.67	0.58	0.60	0.23
Engagingness ↑	0.44	0.22	0.31	0.26	0.40

Table 6. Standard deviations between Turker's rating with respect to Table 2.

Category	**Setting**				
	BB3 Original	**Our BB3 Version**	**BB3 No Memory**	**BB3 No Entity**	**Vanilla BB3**
Repetitions	0.16	0.11	0.17	0.12	0.07
Irrelevancy	0.03	0.04	0.03	0.04	0.05
Contradictions	0.04	0.02	0.07	0.04	0.05
Hallucinations	0.04	0.03	0.04	0.03	0.02

Table 7. Comparison of latency between our implementation of BlenderBot vs. original BlenderBot and Vanilla BlenderBot (i.e., BB3 with only its generative module). Experiments were conducted on Nvidia Tesla T4 GPUs and measured with conversational contexts obtained during data collection.

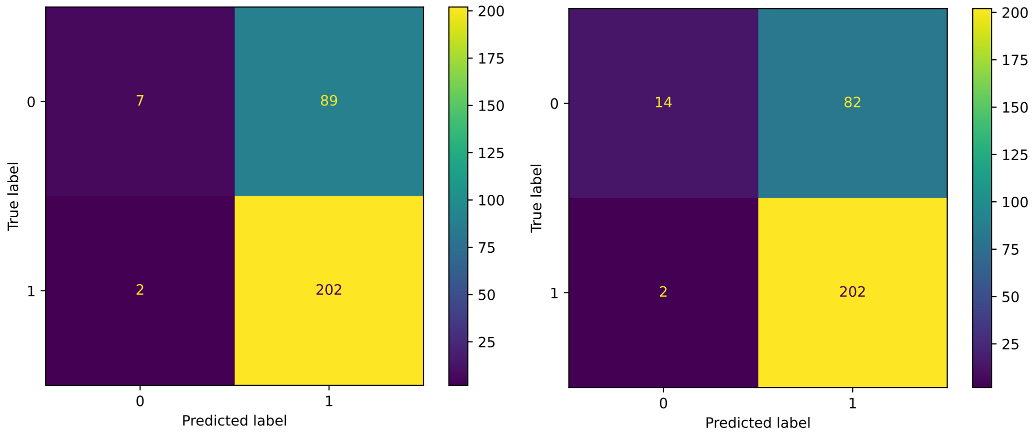
Category	**Setting**			
	BB3 Original	**Our BB3 Version**	**Vanilla BB3**	**BB3 with APIHUB**
Speedup	1x	3.3x	7.2x	1.2x
Latency (mean)	5.1 s	1.5 s	0.7 s	4.6 s



(**a**) Memory decision: BB3 (eval set)

(**b**) Memory decision: Deberta xsmall (eval set)

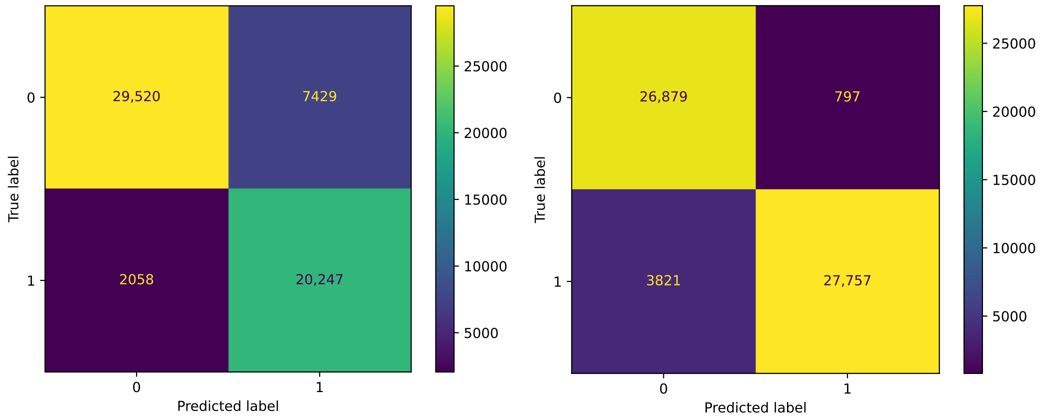
Figure 3. Memory decision results on eval set.



(**a**) Memory decision: BB3 (test set)

(**b**) Memory decision: Deberta xsmall (test set)

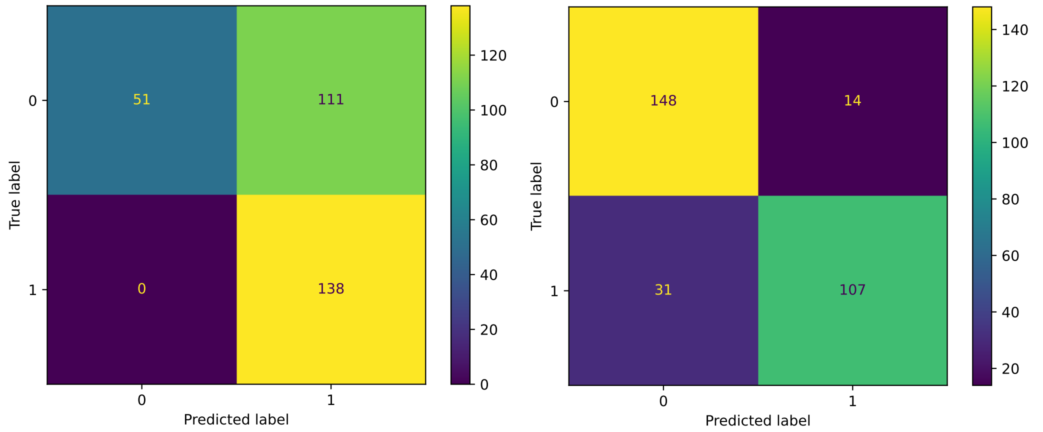
Figure 4. Memory decision results on test set.



(**a**) Search decision: BB3 (eval set)

(**b**) Search decision: Deberta xsmall (eval set)

Figure 5. Search decision results on eval set.



(**a**) Search decision: BB3 (test set)

(**b**) Search decision: Deberta xsmall (test set)

Figure 6. Search decision results on test set.

6. Conclusions

Our research thoroughly scrutinizes and enhances the BlenderBot 3 generative chat model and proposes a chatbot framework that (although it runs by default with the BB3 3B model) can run with any generative model, such as BlenderBot or Llama. We identified and addressed several issues in the original BlenderBot 3, such as long-term memory and entity recognition, resulting in an improved system with superior performance and efficiency.

Our study highlights the importance of dataset quality in task performance and demonstrates that simpler models like DeBERTa-xsmall can outperform complex ones like BlenderBot 3 when fine-tuned with carefully curated datasets. We found that eliminating ineffective processes can streamline the system, decreasing semantic repetitions and computational complexity.

We propose using robust tools (like Amazon EVI, a semantic Wikipedia search, and FiD) for knowledge extraction. We also fine-tuned the model on an expansive conversational dataset using negative sampling, leading to more engaging responses and especially reducing repetitions.

Our enhancements not only boost the chatbot's performance but also streamline its architecture, improving computational efficiency (our modified chatbot outperforms the original BlenderBot 3 by almost 9% in terms of overall quality and is more than three times faster). These advancements lay the groundwork for future generative chat model research. Our work is open-sourced for transparency and shared knowledge.

Author Contributions: Methodology—O.K., J.C., Software—O.K., D.H., J.C., T.G.; Investigation—O.K., Writing—original draft—O.K., J.C., Writing—review & editing—O.K., D.H., Visualization—J.C., Project administration—J.P.; Conceptualization—P.M., Funding acquisition—J.K.; Supervision—J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

Appendix A. Original BlenderBot 3 Architecture

The BB3 system utilizes a solitary generative model to execute all tasks inherent to its architecture. Initially, the system assesses three key decisions: whether to incorporate external knowledge, whether to use its long-term memory, where it archives summaries of user responses in a database as vector embeddings across multiple sessions, and whether to derive a dominant entity, which refers to the primary subject of the current dialogue.

When BB3 opts for the utilization of its long-term memory, it retrieves a summary from the memory database via DPR. In instances where external knowledge is necessitated, BB3 formulates a search query based on the user's input, input that is then routed to their chosen search engine, Mojeek. Mojeek, in turn, produces k relevant documents, from which BB3 applies FiD to distill specific knowledge into a concise form.

The integration of the long-term memory-acquired knowledge and dominant entity with the user's input and the conversational context is carried out within the BB3 generative model, culminating in the generation of the final dialogue. Subsequently, a summary is formulated based on the user's input and BB3's response, which is then stored for future reference in its long-term memory.

Appendix B. Examples of BlenderBot 3's Weaknesses

We conducted tests on BlenderBot 3 to identify systematic flaws in the model. Below, we present examples of the main flaws discovered. Please note that 'Bot' refers to BlenderBot 3, and 'User' refers to a tester, a member of the team.

1. Sometimes it fails to foster deeper conversations, resulting in shallowness. This is associated with another flaw: BB3 sometimes switches topics too quickly.

```
Bot:> What is the best way to discover new music?
User:> I guess the recommender on spotify or youtube
Bot:> Do you have a favorite band?
User:> not just one
Bot:> Do you like to listen to music while you work?
```

2. Repetitions (i.e., utterances with the same semantic as BB3's previous utterance or part of BB3's input).

```
Bot:> Do you believe that there is life
      in our solar system?
User:> yeah, I believe so
Bot:> Do you believe in extraterrestrial life?
User:> yes
Bot:> Do you have any other questions?
User:> what do you believe in?
Bot:> I believe in the existence
      of extraterrestrial civilizations.
```

3. Non-relevant and nonsensical outputs (given input).

```
Bot:> When was the last time you stopped
      to appreciate something beautiful?
User:> today
Bot:> I went to St. Louis.
User:> and?
Bot:> St.Louis is the largest city in Illinois.
User:> and what about it?
Bot:> St.Louis is the 2nd largest city of Missouri.
```

4. Succinct outputs.

```
Bot:> I like chess. I play Xiangqi chess.
User:> is it some special sort of the chess game?
Bot:> Alright.
```

5. Hallucination or false statements.

```
Bot:> St.Louis is the largest city in Illinois.
```

6. Contradictions

```
User:> have you ever been to Vienna?
Bot:> I have not been to the capital of Austria,
      but I have been to the
      largest city on the river Danube.
User:> What is the largest city on the river Danube?
Bot:> I am not sure, but it is in Austria.
User:> is it Vienna?
Bot:> Yes, it is Vienna. It’s the capital
      and largest city of Austria with about
      2 million people.
User:> Have you been there?
Bot:> I have been there. It was a great experience.
```

7. Sometimes it does not understand the user’s question:

```
User:> What do you like to read?
Bot:> I do, I love to read
```

8. High latency.

References

1. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
2. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155.
3. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
4. 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:2208.03188
5. University of California; Barbara, S. GauchoChat: Towards Proactive, Controllable, and Personalized Social Conversation. In Proceedings of the Alexa Prize SocialBot Grand Challenge 5 Proceedings. 2023. Available online: https://www.amazon.science/alexa-prize/proceedings/gauchochat-towards-proactive-controllable-and-personalized-social-conversation (accessed on 12 September 2023).
6. Stevens Institute of Technology. From Hybrid Dialogers to Neural Responders. In Proceedings of the Alexa Prize SocialBot Grand Challenge 5 Proceedings. 2023. Available online: https://www.amazon.science/alexa-prize/proceedings/nam-from-hybrid-dialogers-to-neural-responders (accessed on 12 September 2023).
7. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
8. 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:2004.13637.
9. Komeili, M.; Shuster, K.; Weston, J. Internet-Augmented Dialogue Generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; pp. 8460–8478. [CrossRef]
10. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *arXiv* **2021**, arXiv:2007.01282.
11. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2021**, arXiv:2005.11401.
12. Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv* **2020**, arXiv:2004.04906.
13. Xu, J.; Szlam, A.; Weston, J. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. *arXiv* **2021**, arXiv:2107.07567.
14. Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; Weston, J. Wizard of Wikipedia: Knowledge-Powered Conversational agents. *arXiv* **2019**, arXiv:1811.01241.
15. Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval Augmentation Reduces Hallucination in Conversation. *arXiv* **2021**, arXiv:2104.07567.
16. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]

17. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.01219.
18. Lee, J.; Shim, M.; Son, S.; Park, C.; Kim, Y.; Lim, H. There is no rose without a thorn: Finding weaknesses on BlenderBot 2.0 in terms of Model, Data and User-Centric Approach. *arXiv* **2022**, arXiv:2201.03239.
19. Shuster, K.; Komeili, M.; Adolphs, L.; Roller, S.; Szlam, A.; Weston, J. Language Models that Seek for Knowledge: Modular Search and Generation for Dialogue and Prompt Completion. *arXiv* **2022**, arXiv:2203.13224.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762.
21. 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:1910.13461.
22. Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; Launay, J. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *arXiv* **2023**, arXiv:2306.01116.
23. Stanford University. Dialogue Distillery: Crafting Interpolable, Interpretable, and Introspectable Dialogue from LLMs. In Proceedings of the Alexa Prize SocialBot Grand Challenge 5 Proceedings. 2023. Available online: https://www.amazon.science/alexa-prize/proceedings/chirpy-cardinal-dialogue-distillery-crafting-interpolable-interpretable-and-introspectable-dialogue-from-llms (accessed on 12 September 2023).
24. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Stanford Alpaca: An Instruction-following LLaMA Model. 2023. Available online: https://github.com/tatsu-lab/stanford_alpaca (accessed on 13 March 2023).
25. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling Instruction-Finetuned Language Models. *arXiv* **2022**, arXiv:2210.11416.
26. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2020**, arXiv:1910.10683.
27. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling Language Modeling with Pathways. *arXiv* **2022**, arXiv:2204.02311.
28. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Proceedings of the ACL, Florence, Italy, 28 July 2019.
29. Radlinski, F.; Balog, K.; Byrne, B.; Krishnamoorthi, K. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Stockholm, Sweden, 11–13 September 2019.
30. Henderson, M.; Budzianowski, P.; Casanueva, I.; Coope, S.; Gerz, D.; Kumar, G.; Mrkšić, N.; Spithourakis, G.; Su, P.H.; Vulic, I.; et al. A Repository of Conversational Datasets. In Proceedings of the Workshop on NLP for Conversational AI, July 2019. Available online: https://www.github.com/PolyAI-LDN/conversational-datasets (accessed on 16 April 2019).
31. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv* **2019**, arXiv:1804.07461.
32. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv* **2020**, arXiv:1909.10351.
33. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv* **2020**, arXiv:2003.10555.
34. Atkins, C.; Zhao, B.Z.H.; Asghar, H.J.; Wood, I.; Kaafar, M.A. Those Aren't Your Memories, They're Somebody Else's: Seeding Misinformation in Chat Bot Memories. *arXiv* **2023**, arXiv:2304.05371.

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.