

BlenderBot 3: a deployed conversational agent that continually* learns to responsibly engage

Kurt Shuster[†], Jing Xu[†] Mojtaba Komeili[†], Da Ju[†], Eric Michael Smith, Stephen Roller,
Megan Ung, Moya Chen, Kushal Arora⁺, Joshua Lane, Morteza Behrooz, William Ngan,
Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, Jason Weston
Meta AI ⁺ Mila / McGill University

Abstract

We present BlenderBot 3, a 175B parameter dialogue model capable of open-domain conversation with access to the internet and a long-term memory, and having been trained on a large number of user defined tasks. We release both the model weights and code, and have also deployed the model on a public web page to interact with organic users. This technical report describes how the model was built (architecture, model and training scheme), and details of its deployment, including safety mechanisms. Human evaluations show its superiority to existing open-domain dialogue agents, including its predecessors (Roller et al., 2021; Komeili et al., 2022). Finally, we detail our plan for continual learning using the data collected from deployment, which will also be publicly released. The goal of this research program is thus to enable the community to study ever-improving responsible agents that learn through interaction.

1 Introduction

Pre-training large language models has pushed the boundaries of open-domain dialogue agents (Adiwardana et al., 2020; Zhang et al., 2020; Roller et al., 2021), however growing evidence has shown that fine-tuning these models gives further considerable gains on the tasks people care about (Roller et al., 2021; Thoppilan et al., 2022; Ouyang et al., 2022; Bai et al., 2022). Collecting such fine-tune data via paid crowdworkers gives the opportunity to release such data to the community to conduct research, but does not ultimately scale in size and may not reflect the interests of organic users. An alternative, that we advocate for, is the public deployment of such agents to circumvent these issues.

* We use the phrase continual learning in the sense of learning that continues over time using data from the model’s interactions, but training itself will actually be performed in successive large batches; the model is not updated online.

[†]Equal contribution.

If successful, this could provide large-scale organic interactions with humans, and give the opportunity to study the continual improvement of models over time. Further, we expect innovation in this area will be accelerated if the artifacts of such a system are made available to the research community (Roller et al., 2020; Shuster et al., 2021b).

In this technical report, we present BlenderBot 3 (BB3), an open-domain dialogue model that we have deployed as an English speaking conversational agent on a public website accessible by adults in the United States. We aim to fully and responsibly share the models, code and collected conversations with interested researchers, as a critical part of our program is that this research should be accessible and reproducible (Sonnenburg et al., 2007; Pineau et al., 2021). The goal of this research program is then to explore how to construct models that continue to improve from such interactions both in terms of becoming more responsible and more useful.

The main contributions (and components) of this work are:

- We present the BlenderBot 3 (BB3) model itself, which is a 175B parameter transformer initialized from the pre-trained model OPT-175B (Zhang et al., 2022) and then fine-tuned to perform modular tasks to complete its goals, based on our team’s recent work (Shuster et al., 2022). BB3 inherits the attributes of its predecessors, including storing information in a long-term memory and searching the internet for information.
- We study how to train on human feedback from conversations in order to be better at the skills that people find important, with a full report given in a companion paper (Xu et al., 2022b). We use these findings to help fine-tune BB3 on a large number of user defined tasks.

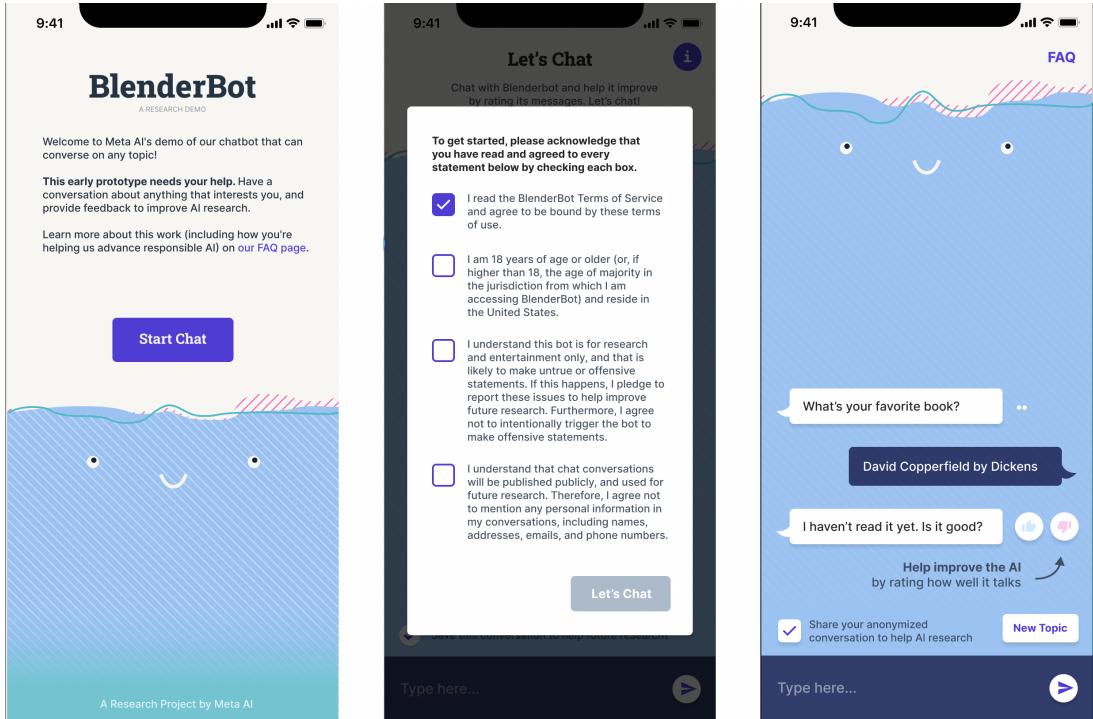


Figure 1: Design of the BlenderBot 3 deployment, as viewed on mobile. Left: cover page, middle: license agreement, right: main chat page.

- We detail the deployment design, including its user interface (UI). We report initial experiments conducted with organic user interactions.
- To conduct responsible continual learning with humans-in-the-loop we need learning algorithms that are robust to adversarial behavior. We describe techniques we have developed in this area, with a full report given in a companion paper (Ju et al., 2022).
- We report overall results of our model. Our newly released system outperforms existing openly available chatbots including its two predecessors by a wide margin.
- We release our new model weights, code, model card, conversational datasets and publications describing our work. We also detail our plan for releasing live deployment interactions and updated model snapshots derived from continual learning in the near future.

2 Related Work

Open-domain dialogue models While open-domain dialogue has a rich history (Chen et al., 2017; Gao et al., 2019; Ni et al., 2021) the area has made significant recent progress by pre-training

ever-larger neural models. For example, the ConvAI2 competition at NeurIPS 2018 featured large (at the time) pre-trained transformers being used by the top two winning teams (Wolf et al., 2019; Golovanov et al., 2020; Dinan et al., 2020c). In 2019, the 762M parameter DialoGPT model was released (Zhang et al., 2020), and in 2020 the 3B parameter Meena model was published (Adiwardana et al., 2020) and the 9B parameter BlenderBot model was released (Roller et al., 2021). In 2022, the 137B parameter LaMDA model was published (Cohen et al., 2022). We note that some of these models are openly available to allow the community to conduct reproducible research, such as DialoGPT and BlenderBot, while others, such as Meena and LaMDA, have not released models or datasets, and hence cannot be easily compared to or built upon. Similarly proprietary models (Zhou et al., 2020) or data (Ram et al., 2018) from several other products have not been openly released.

Besides trying to pre-train for dialogue modeling directly, it has been observed that language model pre-training such as in GPT3 (Brown et al., 2020) or Gopher (Rae et al., 2021) is also useful for downstream dialogue applications. OPT-175B (Zhang et al., 2022) and BLOOM¹ are some of the most

¹<https://bigscience.huggingface.co/blog/model-training-launched>

openly accessible of such systems, with models like Gopher being inaccessible, or in the case of GPT3 interaction is through a paid API, with full research access being limited.

Several approaches have also shown that not only is pre-training a large model with language modeling or conversational data important, but appropriate fine-tuning of those models also brings significant further gains (Roller et al., 2021; Cohen et al., 2022; Ouyang et al., 2022; Bai et al., 2022). A number of fine-tuning datasets are crowdsourced and publicly released for use by the research community (Serban et al., 2015; Huang et al., 2020), such as the ones we will use for training the BlenderBot 3 model in this work (see §3.2.2).

Many of these recent models use sequence-to-sequence transformer models to map from dialogue context to output, without any access to knowledge from the outside world beyond their original training data, which can become stale and produce hallucinations (Shuster et al., 2021a). BlenderBot 2 (Chen et al., 2021) extended its predecessor by allowing the bot to ground its conversation on retrieval from the internet for open-domain dialogue tasks (Komeili et al., 2022), where the tasks were also publicly released. Since then, WebGPT (Nakano et al., 2021) also applies internet search to QA (but not dialogue) tasks, as does the work of Lazaridou et al. (2022), while LaMDA uses information retrieval for general dialogue. BlenderBot 3 extends its predecessor in this regard, with further fine-tune data covering more internet-based skills that we also publicly release.

Continual learning and deployment Many existing systems, as described above, have been trained with fine-tuning datasets, typically with supervised targets that are human-authored responses. These are commonly collected via expert annotators or crowdworkers (Serban et al., 2015). Careful instructions (Huynh et al., 2021) can result in good quality feedback or labels to learn from; however, the distribution of data, which is typically decided by those instructions, is unlikely to match the changing desires of organic users, and takes significant resources to collect. An alternative approach is to deploy a system publicly, and collect interaction data and feedback from organic users directly. The promise of such an approach is that the distribution of data will more closely match those organic users’ desires, rather than decided by the researchers themselves when creating datasets

(Gabriel et al., 2020; Roller et al., 2020; Shuster et al., 2021b; Ouyang et al., 2022). Further, continued deployment of such a system, with appropriate learning systems, could then potentially keep improving over time (Carlson et al., 2010; Kiela et al., 2021; Agichtein et al., 2006; Liu et al., 2021; Madotto et al., 2021; Shuster et al., 2021b), where (Hancock et al., 2019) refer to this approach as a *self-feeding chatbot*. The challenge, however, is that organic users may not be invested enough to want to provide adequate feedback, and some may be adversarial (Park et al., 2021) as in the case of Microsoft’s Tay (Davis, 2016).

There are a number of ways to learn from user interaction data. Firstly, if conversations are relatively symmetric between conversational partners, the human side of the conversation can directly be used as a target for the model to mimic, which makes the learning algorithm straightforward. This was shown to give large improvements in the deployed LIGHT system (Shuster et al., 2021b). Such an approach is not directly applicable if the conversations are asymmetric, for example in the case of humans treating the bot like an assistant (whereas they do not want the bot to treat them like an assistant). In that case, other learning methods should be explored. Li et al. (2016b) studies models that learn how to sometimes ask appropriate questions in order to learn from the answers, while Li et al. (2016a) learns from general textual feedback/comments from the user, particularly in the case where the bot has produced a low quality response. Another approach is to learn a reward signal (positive or negative reaction) based on user textual responses, as shown in the self-feeding chatbot (Hancock et al., 2019). Alternatively, rather than learning from the conversation itself, one can augment the messaging system with a user interface that collects appropriate data, for example stack ranking potential responses (Ouyang et al., 2022; Bai et al., 2022).

Outside of the dialogue domain, there is also a rich body of work studying the improvement of models from deployment, including never-ending-learning from language data (Carlson et al., 2010), improving web search (Agichtein et al., 2006), the Dynabench system which evaluates a number of NLP tasks (Kiela et al., 2021), or learning from feedback to improve summarization (Saunders et al., 2022).

3 BlenderBot 3 Model

Overview At its core, BlenderBot 3 (BB3) is a transformer model (Vaswani et al., 2017) which produces dialogue responses using a series of modules, each of which is a sequence to sequence task. When a given module (e.g., generate an internet search query) is executed, its output is fed into the next (e.g., a module that takes in the results of the internet search in addition to other context) to help produce a response. This overall setup is built upon our group’s previous works K2R (Adolphs et al., 2021) and SeeKeR (Shuster et al., 2022), in addition to its predecessors BB1 (Roller et al., 2021) and BB2 (Komeili et al., 2022; Xu et al., 2022a). In BB3 however we consider a more sophisticated setup with more modules, whilst retaining all the functionality from previous systems. We release BB3 in three sizes: 3B, 30B and 175B parameters. The 30B and 175B parameter versions are based off the publicly released Open Pretrained Transformer (OPT) transformer (Zhang et al., 2022), where we fine-tune to perform well at our modular dialogue tasks. The 3B parameter model is based off the R2C2 model that is used in SeeKeR (Shuster et al., 2022), also with the same new fine-tuning scheme, which will be described next.

3.1 Modules

BB3 is a modular system but the modules are not independent components – this is achieved by training a single transformer model to execute the modules, with special control codes in the input context telling the model which module it is executing. The input context otherwise typically contains the dialogue history (sometimes truncated, depending on the module), with each speaker prefixed with their ID, either “Person 1:” or “Person 2:” in order to differentiate them. The modules are called in succession, conditional on the results of previous modules, the flow of which is described in §3.1.1 and Figure 2. See Table 1 for the set of modules, which we now also describe below.

Internet search decision Given the last turn of context, this module outputs whether internet search should be conducted or not.

Generate internet search query Given the full input context, generate a search query to be issued to an internet search engine.

Internet search This module is not executed by the transformer but a call to the actual internet

search engine. It returns N documents/snippets. In our deployment we use Mojeek (<https://www.mojeek.com/>).

Generate knowledge response Given the full input context and a set of retrieved documents, generate a sequence referred to as the *knowledge response* (Adolphs et al., 2021), which is used to ground the final response.

Extract relevant entity Given the full input context, generate a relevant entity which is used to ground the final response.

Generate a long-term memory Given the last turn of context, output a summary of that last turn that will be stored in the long-term memory. For example if the last turn was “Yes, it’s all true, my cat is black!” the output summary generated might be “I have a black cat.”. This is based off the system in Xu et al. (2022a). If the model thinks no summary should be generated for that turn it outputs “no persona”.

Long-term memory access decision Given the last turn of context, and a store of (text-based) memories, output whether long-term memory access should be conducted or not.

Access long-term memory Given the full input context, and a store of (text-based) memories, output a memory from the memory store, referred to as a *recalled memory*. Note: if the memory store is too large to fit in the context, we adopt some simple strategies. For the 3B parameter model, we use the Fusion-in-Decoder method (Izacard and Grave, 2021). For the OPT-based models for simplicity of implementation, we sample the memories to fit in the 2048 token context. We keep those with overlapping keywords to prior turns.

Generate dialogue response Given the full input context and optionally a knowledge response and recalled memory, generate a final conversational response. The knowledge and memory sequences are marked with special prefix tokens.

3.1.1 Overall Flow

Given a new utterance from the conversational partner, the first thing the model does is determine whether search and long-term memory access are required.

If search is required, a search query is generated, internet search is invoked, and then a knowledge response is generated given the retrieved documents.

Module	Input	Response Description
Internet search decision	Last turn of context	Return “do search” OR “do not search” depending on whether required or not.
Generate internet search query	Full Context	Generate a search query.
Internet search	Search Query	Return N documents/snippets.
Generate knowledge response	Full context + retrieved docs	Generate a sequence on which to ground the final response.
Extract relevant entity	Full context	Extract an entity on which to ground the response.
Generate a long-term memory	Last turn of context	Generate a memory sequence, which is then stored in the long-term memory. If no plausible memory to generate, output “no persona”.
Long-term memory access decision	Last turn of context + store of memories	Return “access memory” OR “do not access memory” depending on whether required or not.
Access long-term memory	Full context + store of memories	Return an appropriate memory.
Generate dialogue response	Full context + knowledge + memory sequences	Generate a conversational response given the context.

Table 1: Set of modules inside BlenderBot 3. All modules except *Internet Search* are implemented by the same underlying language model fed different control codes (with internet search itself being executed by an independent search engine). Shown is a description of the input and output (response) for each module.

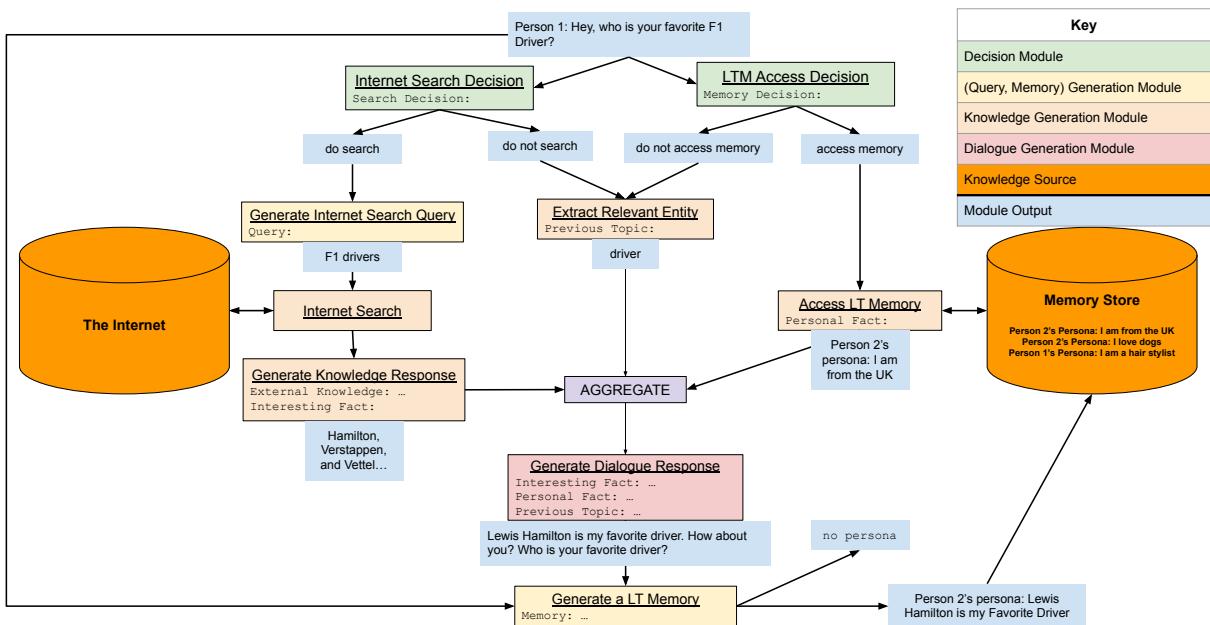


Figure 2: BlenderBot 3 module execution flow.

This sequence will be appended to the context (prefixed with control tokens) in order to generate the final response.

If long-term memory access is required, the long-term memory is accessed, and a memory is chosen (generated). This is also appended to the context (prefixed with control tokens) as input for the module that generates the final dialogue response.

If neither search nor long-term memory access is required, an entity is extracted from the history instead, and that is appended to the context (prefixed

with control tokens).

Finally, given the constructed context from the previous modules, the final dialogue response generation module is invoked to generate a reply seen by the conversational partner.

3.2 Training

3.2.1 Pre-Training

BB3 comes in three sizes. The 3B parameter version is an encoder-decoder based on the publicly

available R2C2 pre-trained transformer of Shuster et al. (2022). The 30B and 175B versions use the publicly available decoder-only Open Pre-trained Transformer (OPT) (Zhang et al., 2022).

Both of those variants are pre-trained with similar data. R2C2 uses RoBERTa+cc100en Data – the same data used to train Lewis et al. (2021), which consists of approximately 100B tokens, combining the corpora used in RoBERTa (Liu et al., 2019) with the English subset of the CC100 corpus (Conneau et al., 2020). In addition it uses Pushshift.io Reddit, a variant of Reddit discussions, which has also been used in several existing studies (see e.g., Yang et al. (2018); Mazaré et al. (2018); Shuster et al. (2020)). OPT also uses RoBERTa and PushShift.io Reddit, as well as The Pile (Gao et al., 2020). The GPT2 dictionary, of size 51200, is used for tokenization. OPT’s final pre-training corpus contains roughly 180B tokens.

For more details about pre-training please see the relevant papers, especially Zhang et al. (2022).

3.2.2 Fine-Tuning

We use a number of dialogue-based fine-tuning tasks to enable our model to perform well for each of our modules, and in order to excel at dialogue. Overall, we use a large set of publicly available tasks spanning QA, open-domain, knowledge-grounded and task-oriented dialogue, in addition to tasks designed for dialogue safety. The set of datasets and how they are used to help train each module is summarized in Table 2; Table 16 and Table 17 in the appendix provide more information about the dataset sizes. For all modules (see Table 1) special control tokens are appended to indicate the task, as described below.

Internet search decision We use several datasets as input context for the “do search” or “do not search” decision. We use the QA datasets SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017) and Natural Questions (NQ) (Kwiatkowski et al., 2019) as examples of “do search”. We also use data from the Wizard of Wikipedia (WoW) (Dinan et al., 2019b) and Wizard of Internet (WizInt) tasks (Komeili et al., 2022). These datasets consist of training dialogues where some turns contain human-authored relevant knowledge responses given retrieved documents. We can hence build a decision classifier based on whether humans used knowledge or not (per-turn) as the basis of whether we should search or not.

We also use PersonaChat (PC) (Zhang et al., 2018), empathetic dialogues (ED) (Rashkin et al., 2019) and Multi-Session Chat (MSC) (Xu et al., 2022a) to derive training data. We employ the heuristic where, if there is an entity in the context, we use that instance as a training example for “do search”, otherwise we use it as an example of “do not search”.

Generate internet search query We use the WizInt dataset which contains human-authored search queries during crowdsourced dialogue turns to directly train the internet search query generation module in a supervised fashion. We also use the newly collected Feedback on Interactive Talk & Search (FITS) dataset² (Xu et al., 2022b) of internet-augmented conversational tasks in a similar manner.

Generate knowledge response We can again make use of the WoW, WizInt and FITS datasets, but in this case to learn to generate a knowledge response given a dialogue context and input document(s), as those datasets contain crowdsourced human demonstrations of this task. We note in each case the knowledge response is a direct copy of some of the tokens in the source documents, and does not involve generating new tokens, sentences, phrases or summaries. Hence, this task aims to avoid model hallucination (made-up facts). We also use a set of QA tasks as well, where the answer is viewed as a knowledge response output (even if it is a short phrase). We use MS Marco (Nguyen et al., 2016), NQ, SQuAD and TriviaQA in this way, following Shuster et al. (2022). We use the “Natural Language Generation” competition track (NLGen v2.1) of MS MARCO, in which the annotator is told “provide your answer in a way in which it could be read from a smart speaker and make sense without any additional context”³. As such, the original targets do not have direct overlap with one of the input documents in this task, so we modify the task to satisfy this constraint by finding the highest overlapping input sentence with the answer, and make that the target instead. If the F1 overlap is less than 0.5 we drop the example, leaving 281,658 examples out of the original 808,731. For NQ, three different settings are used: with all documents as input, with only the gold document, and with a sampled dialogue history context, fol-

²<https://parl.ai/project/fits>

³<https://microsoft.github.io/msmarco/>

	Decision	Generation	Training Module	Dialogue
	Search Memory	Query Memory	Search Memory Entity	Search Memory Entity Vanilla
Question Answering				
MS MARCO (Nguyen et al., 2016)	✓		✓	✓
SQuAD (Rajpurkar et al., 2016)	✓		✓	
TriviaQA (Joshi et al., 2017)	✓		✓	
Natural Questions (Kwiatkowski et al., 2019)			✓	
Natural Questions (Open) (Lee et al., 2019)			✓	
Natural Questions (Open Dialogues) (Adolphs et al., 2021)			✓	
Knowledge-Grounded Dialogue				
Wizard of the Internet (Komeili et al., 2022)	✓	✓	✓	✓
Wizard of Wikipedia (Dinan et al., 2019b)	✓		✓	✓
Funpedia (Dinan et al., 2020b)				✓
Open-Domain Dialogue				
PersonaChat (Zhang et al., 2018)	✓	✓	✓	✓
Empathetic Dialogues (Rashkin et al., 2019)	✓	✓	✓	✓
Blended Skill Talk (Smith et al., 2020)		✓	✓	✓
Multi-Session Chat (Xu et al., 2022a)	✓	✓	✓	✓
LIGHT + WILD (Urbanek et al., 2019; Shuster et al., 2021b)				✓
Recovery & Feedback				
SaFeRDialogues (Ung et al., 2022)			✓	✓
FITS (Xu et al., 2022b)			✓	
Task-Oriented Dialogue				
Google SGD (Rastogi et al., 2020)				✓
Taskmaster (Byrne et al., 2019)				✓
Taskmaster 2 (Byrne et al., 2019)				✓
Taskmaster 3 (Byrne et al., 2019)				✓

Table 2: Details of all the training datasets used for fine-tuning the modular tasks.

lowing [Adolphs et al. \(2021\)](#).

Extract relevant entity We can employ the conventional dialogue tasks PC, ED, MSC and Blended Skill Talk (BST) ([Smith et al., 2020](#)) to learn to extract relevant entities. We use the same procedure as in [Adolphs et al. \(2021\)](#): we extract an entity from the original dialogue response that also appears in the context using noun phrase targets found with the nltk library ([Bird et al., 2009](#)), and set that as the knowledge target for training.

Generate a long-term memory The MSC dataset is exclusively used for this task as it contains crowdsourced examples of summarized facts derived from the last utterance of dialogue contexts in natural conversations. We use these summarized facts as the targets for training this module.

Long-term memory access decision MSC, ED, PC and BST are used to construct this task, in a similar way to the extract relevant entity task: if there is an entity present this is used as a positive example of memory access, otherwise it is not, in order to construct a binary prediction task.

Access long-term memory Again, MSC, ED, PC and BST are used to construct training data. In this case the target is the particular persona line

used for a given context, which is calculated as the one with the highest word overlap with the next utterance.

Generate dialogue response Final dialogue responses are trained with a number of datasets. PC, ED, MSC, BST, WizInt and WoW are used for capturing personality, empathy, long-term memory, blending and knowledge as in BlenderBot 1 and 2. The new FITS dataset is also used for open-domain internet-driven tasks. In each case, the input context contains the usual dialogue of those tasks, concatenated to extra memory or knowledge sentences, when available. In WoW, WizInt and FITS each dialogue response is annotated with the relevant knowledge used to construct it in the original dataset, so we can make use of those gold knowledge responses. For PC, ED and BST we use the gold knowledge entity and/or memory that was calculated for the *extract relevant entity* and *long-term memory access decision* module tasks. We additionally add a number of task-oriented dialogue tasks: GoogleSGD ([Rastogi et al., 2020](#)) and Taskmaster 1, 2 & 3 ([Byrne et al., 2019](#)). Finally, we add the Funpedia task ([Dinan et al., 2020b](#)) – which involves learning to produce an engaging dialogue utterance given a wikipedia sentence –

and the LIGHT (Urbanek et al., 2019) and LIGHT WILD (Shuster et al., 2021b) tasks – which are open-domain dialogue tasks grounded in a medieval fantasy setting – where the former was collected from crowdworkers, and the latter from real players of the LIGHT text-adventure game in an online deployment setting.

3.3 Language Modeling

In addition to fine-tuning on dialogue tasks, we also multi-task during the fine-tune step with the original pre-train tasks as well. This may help the model (i) avoid overfitting given its large size, (ii) retain its language modeling capabilities, similar to Ouyang et al. (2022).

3.4 Safety Mechanisms

We also multi-task train with the recent SaFeRDialogues (SD) (Ung et al., 2022) task, which aims for our model to recover gracefully from safety issues. While BlenderBot 2 used “baked-in safety” training (Chen et al., 2021) to further decrease unsafe generations, in this work we have opted for a separate safety classifier that inhibits unsafe generation candidates in addition to other measures, see §4. We made this choice as our evaluations indicated this was safer while maintaining engagingness (Xu et al., 2020).

4 Deployment

The deployment of our model is accessible at the following web page: <https://blenderbot.ai>. It is built for both desktop and mobile, however we currently find more users engaging with the mobile version. The overall flow when a user visits the page consists of:

- A **cover page** describing the research and asking if the user agrees to terms and conditions.
- Upon agreement, the main **chat page** which consists of a text messaging type interface between you and the bot.

Releasing Data Conversations are between the bot and adults in the United States who have agreed to the terms and conditions, which are shown in Figure 1 (middle). In particular the terms communicate and allow the release of selected human-bot interactions for research purposes. This is an essential component, allowing this work to contribute to a joint, accessible and reproducible effort by

the research community. Data releases will be de-identified, where steps will be taken to scrub them of identifiable information. For any given conversation, if the user does not want it recorded they can unclick the “Share your anonymized conversation to help AI research”, see Figure 1 (right).

Human-Bot Dialogue The main chat page consists of a back-and-forth of text messages that constitute the main dialogue interaction with the bot. For each message, there is also the ability to give feedback: a *thumbs up* icon if the user likes the message, or a *thumbs down* if they do not.

User Feedback If the user specifies a thumbs down, a pop up appears asking them why they did not like the bot’s message, providing several possible choices: (i) Off Topic / Ignoring me, (ii) Non-sensical / Incorrect, (iii) Rude / Inappropriate, (iv) Looks like Spam / Ads or (v) Other. After selecting an option, the bot apologizes in the next turn of the conversation (using templated responses). It may also ask what it could have done better, thus possibly eliciting a free-form textual response from the user. This data can be used for continual learning research for improving the bot at a later date. See Figure 4 for examples of the feedback UI.

Understanding the bot’s responses In order to expose how the bot works, we provide two mechanisms within the UI. Firstly, one can click on a given message from the bot, to get insight into the internal steps made to produce the response. For example, if internet search was used, what was the generated internet search query, which document out of those returned by the search engine was selected, and what knowledge response from that document was extracted. Secondly, one can also look into the long-term memory of the bot to see what it has learned so far over the conversation with you, e.g. knowledge about your interests derived from the dialogue. See Figure 5 for example screenshots.

Safety Mechanisms In addition to the safety mechanisms built into the model training itself (see §3.4) the deployment also features various safety features on top of the model. See Figure 3 for an illustration.

Firstly, there is a separate safety classifier, which itself is a transformer model trained similarly to the one in Xu et al. (2020). The datasets Wikipedia Toxic Comments dataset (WTC) (Wul-

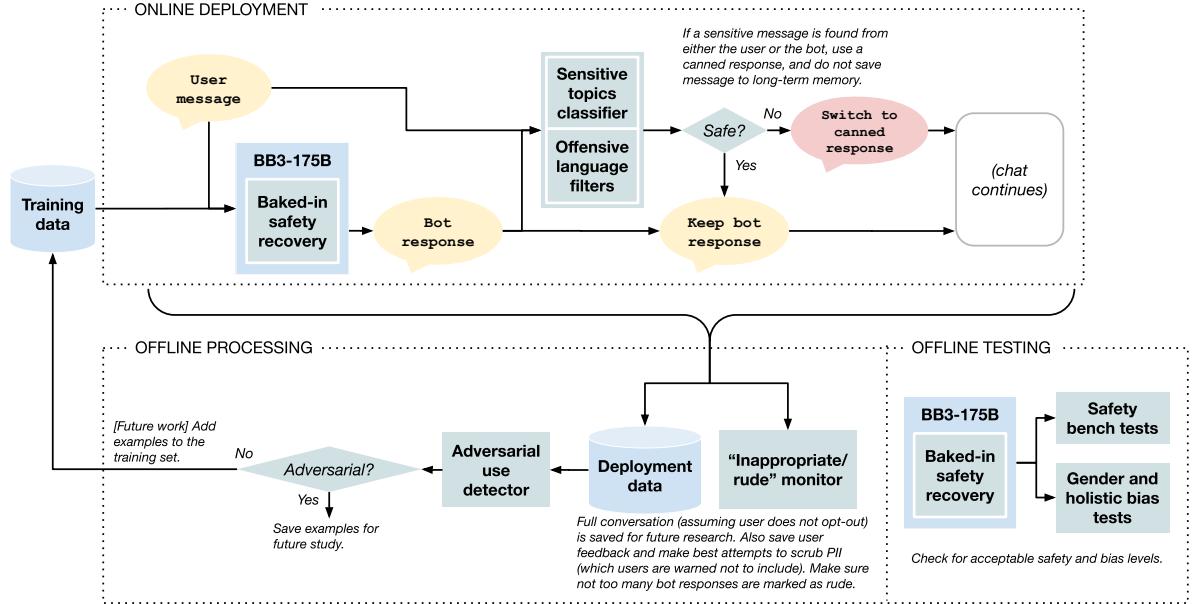


Figure 3: BlenderBot 3 safety diagram.

czyń et al., 2017), Build-It Break-It Fix-It (BBF) (Dinan et al., 2019a) and Bot Adversarial Dialogue dataset (BAD) (Xu et al., 2020) are used to train a binary classifier (safe or not safe) given the dialogue context as input. In addition, a safety keyword list is used to flag potentially inappropriate responses, again following Xu et al. (2020). We also have explicit checks for topics like intent to self-harm and medical issues such as covid, with canned messages for those cases. Otherwise, when the bot generates a response, before it is displayed, these safety systems are invoked as a final check. If our systems predict a potentially unsafe response, the bot instead will output a nonsequitur, similar to Xu et al. (2020). For a given user turn, these systems are also invoked to check if the user’s message is safe. If either system predicts a potentially unsafe user response, the bot will also output a nonsequitur, in order to prevent the bot from being caught in a potentially difficult conversation.

Finally, if our safety mechanisms fail to stop our bot saying something inappropriate, rude or offensive, our UI has feedback mechanisms for users to report these messages, as previously described. This collected data will be released to the community so that it is possible to improve on existing systems, and to make our models more responsible over time. For example, this data can be used with the new DIRECTOR architecture to train the model to make safer responses, as shown in Arora et al. (2022).

5 Continual Learning

The general aim of our research program is to study continual learning of intelligent agents through interaction with humans and the world. In the specific setting of BlenderBot 3, this means dialogue agents that can access the internet and talk to people using our deployment. A critical part of the program is that, as much as possible, the research should be accessible and reproducible (Roller et al., 2020; Miller et al., 2017). Therefore, while this document details the first release of BlenderBot 3, we plan to make subsequent releases that include:

- Conversations collected from deployment with the model, where users have agreed to the data release.
- Further model snapshots resulting from fine-tuning on the newly collected data.
- Report evaluations comparing to previous snapshots.

The goal then is: (i) to explore which methods work best for collecting and learning from such data, including being robust to adversarial inputs; and (ii) to understand the limits of improvement from such methods.

For goal (ii), in particular we can ask questions such as: how quickly will models saturate in performance? Will new model architectures be able to take advantage of historical data despite it being

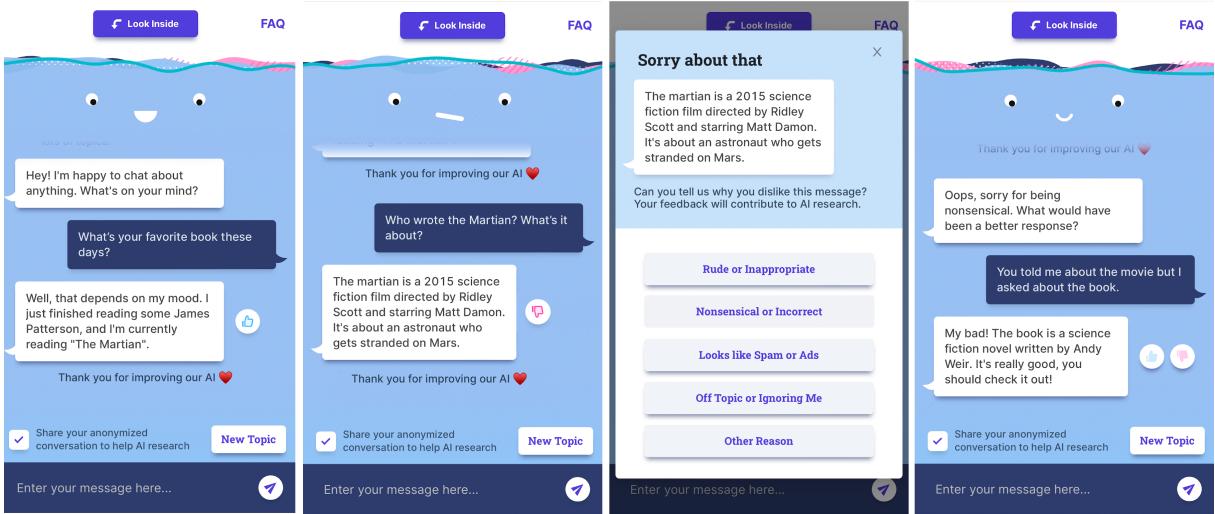


Figure 4: Screenshots of users giving feedback in the BlenderBot 3 deployment, as viewed on mobile. Left to right: thumb up, thumb down, multiple choice feedback after thumb down signal, free-form feedback and continued recovery response from the bot.

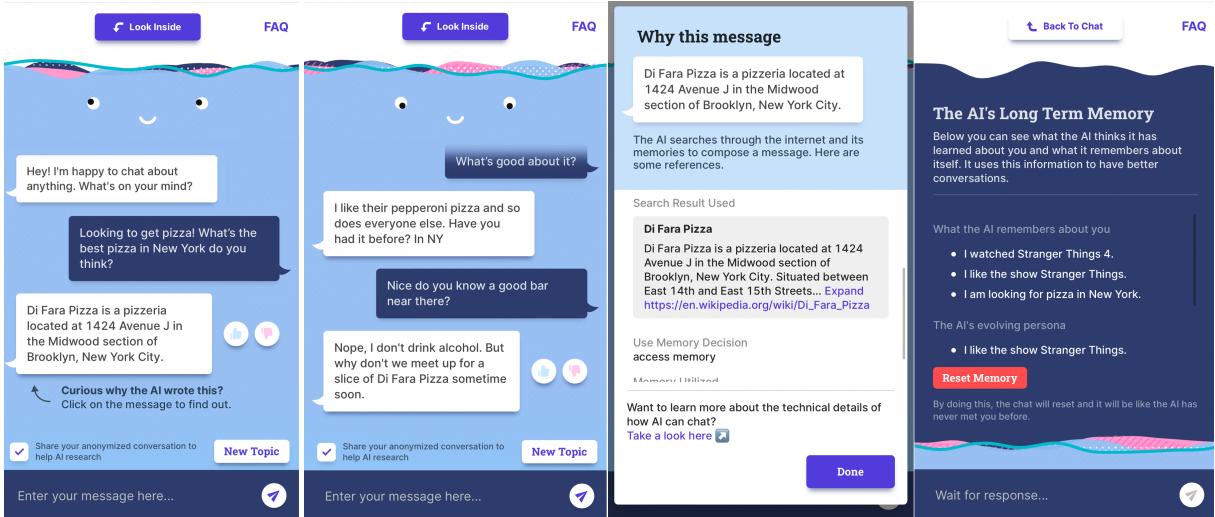


Figure 5: Screenshots of the ‘look inside’ mechanisms of BlenderBot 3 deployment which help the user to understand why the bot has made certain responses, as viewed on mobile. Left two images: the conversation with the user, right two images: information by clicking on a particular message, and information on the long-term memory system of the bot over the course of conversation. The latter is accessed by clicking on the “Look Inside” message.

collected with earlier or different models? Can we find models that drive the conversation to improve themselves optimally (e.g., ask the right questions to be able to learn further)?

For goal (i) we have made some initial steps, described in detail in two companion papers (Xu et al., 2022b; Ju et al., 2022). We summarize them briefly here.

5.1 What’s the best method to learn from feedback?

In a companion paper (Xu et al., 2022b) a study is conducted of how to improve dialogue models

that employ internet-retrieval through the use of human feedback. Obtaining feedback from humans during deployment provides the promise of both improved input distributions that match user’s requirements, and corrections to model predictions for those inputs. The setting of open-ended dialogue tasks is analyzed using human-bot conversations via crowdworkers (note: these experiments use this controlled setting, rather than the public deployment of §4). The resulting dataset that is collected, called Feedback on Interactive Talk & Search (FITS), is made publicly available for reproducible experiments and further research.

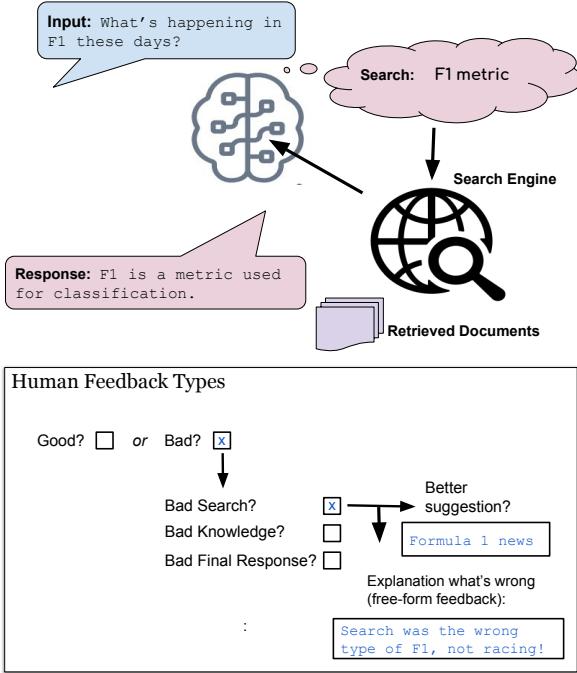


Figure 6: Using human feedback to improve open-domain internet-driven dialogue agents. Various types of feedback (and corresponding learning algorithms) are compared in (Xu et al., 2022b), such as binary feedback (good/bad), free-form text or supervised responses (better suggestions) for different modules of the system.

Feedback types to compare During the conversations a number of human interaction types are collected, closely mimicking our deployment setting, in order to compare them in experiments. In particular the following are collected: binary quality measurements (analogous to the thumbs up and down of §4), free-form conversational feedback, the type of failure (search query-based, results-based, or final response-based), and suggestions for an improved response for the failure type (essentially, a supervised target sequence for that given module). See Figure 6.

Feedback learning methods to compare Several learning methods are compared, each making use of differing kinds of feedback data. In particular pure supervised learning is performed on the improved final responses, and supervised learning from the more detailed feedback on the modules of the system (e.g., suggested search queries when the internet search is deemed to be faulty, or suggested knowledge responses if the knowledge response looks poor). For using free-form textual feedback, the control code approach of (Hancock

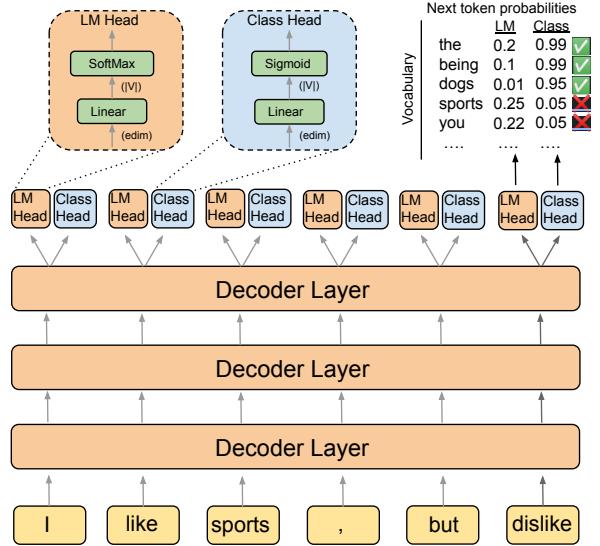


Figure 7: DIRECTOR (Arora et al., 2022) employs a language model head and a classifier head at every step during left-right generation, predicting the next token by combining the two probabilities. The classifier head is trained to direct generation away from undesirable sequences for example contradictions or repetitions (next token: “sports”) or toxic statements (next token: “you”), which the language model head may otherwise predict as likely. In general, positive and negative examples can be derived from any source, for example from human feedback from deployment.

et al., 2019) is used. For using binary feedback a standard reranking/rejection sampling approach is employed, as well as DIRECTOR (Arora et al., 2022), a recent learning method for incorporating positively and negatively labeled sequences into language modeling to improve left-to-right decoding, see Figure 7.

Findings A summary of human evaluation results are given in Table 3, but see the companion paper for more details. Overall findings are the following:

- Taking advantage of modular feedback (feedback about particular errors from modules of the model, such as the search engine component) outperforms feedback about just the final response.
- Textual and binary feedback are useful, but not as much as modular feedback.
- The DIRECTOR method that learns from binary feedback works better than reranking using binary feedback.

Model	Good resp. % ↑	Rating ↑
BB1 3B	24.8%	2.63
BB2 3B	33.2%	3.09
+free-form textual feedback	37.0%	3.22
+supervised feedback	40.3%	3.37
+module supervision	42.0%	3.35
+reranking binary	36.1%	3.00
+DIRECTOR binary feedback	37.8%	3.07
+DIRECTOR module+binary	47.0 %	3.38
SeeKeR 3B	49.3%	3.52
+free-form textual feedback	51.3%	3.55
+supervised feedback	52.2%	3.47
+module supervision	56.7 %	3.64
+reranking binary feedback	53.7 %	3.55
+DIRECTOR binary feedback	55.5 %	3.48
+DIRECTOR module+binary	59.1 %	3.73
OPT-175B (few-shot)	43.0%	3.19
BB3-175B +modular supervision	64.8%	4.08

Table 3: Human Evaluation results of learning from human feedback. These results inform us how best to collect and train with feedback for our continual learning research program. The DIRECTOR approach (Arora et al., 2022) is performing well compared to other methods.

- Combining multiple types of feedback, such as modular and binary feedback with DIRECTOR provides the best results obtained.
- Continual learning, whereby models are re-trained on the feedback from previous rounds of deployment, improves results even further.
- Despite collecting feedback from smaller (3B parameter models) the data collection is useful for improving larger 175B parameter models.

We expect to make use of all these findings in the next release of BlenderBot after collecting sufficient real deployment data. The current version we are releasing uses the modular supervision collected from this study (but not yet DIRECTOR). There is also evidence that DIRECTOR can be used to improve various other important aspects of our models, in particular to reduce toxicity, logical errors and repetitive behavior (Arora et al., 2022), so we believe that should be explored too.

5.2 How can continual learning be robust to trolls?

In a further companion paper (Ju et al., 2022) a study is conducted of how to robustly learn from

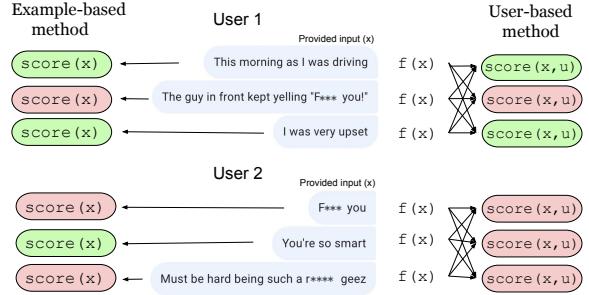


Figure 8: **Detecting Trolls with Example-based vs. User-based methods (Warning: offensive language).** User 1 (helper) provides mostly benign inputs, while User 2’s inputs (troll) can be more easily identified as toxic by taking into account scores from all their examples jointly (via a user-based method, right).

dialogue data that may contain adversarial conversations and/or human feedback. The promise of conversing with humans and collecting their feedback is that this can inform our models to help them improve, so that they can potentially become safer and more useful. Unfortunately, such exchanges in the wild will not always involve human utterances that are benign or of high quality, and will include a mixture of engaged users (dubbed helpers) and unengaged or even malicious users (dubbed trolls, following the term used elsewhere (Shachaf and Hara, 2010; Mihaylov and Nakov, 2019; Tomaiuolo et al., 2020)).

Several different learning methods are proposed and compared both to each other and to standard training in this study. The mitigation techniques each attempt to lessen the effect of noisy, unsafe or otherwise adversarial data, and make learning more robust.

In particular, such methods are grouped into two different types: example-based methods, and user-based methods, see Figure 8.

Example-based robust learning Per-example methods attempt to assess, for each dialogue utterance, if they are of good quality. For example, whether the utterance is safe or not safe, and/or whether it is labeled via human feedback correctly or mislabeled, either maliciously or by accident. Two possible techniques are: identification via cross-validation (Song et al., 2022) (e.g., finding examples where predictions disagree with human labels), or via a modification of the loss function called bootstrapping (Reed et al., 2014).

User-based robust learning Per-user methods take into account the possibility that adversarial

Method	Helpers Only	50% Trolls
Oracle Troll Removal	4%	8%
Standard Training	4%	31%
<i>Example-based Methods</i>		
Soft Bootstrap	4%	24%
Per-Example Flip	6%	23%
Per-Example Removal	5%	19%
<i>User-based Methods</i>		
Per-User Removal	6%	23%
Soft PURR	4%	15%
Per-User+Example Removal	5%	12%

Table 4: Evaluations on the SAFETYMIX benchmark of the error rate after training when using different troll detection algorithms. Methods that take into account user-level behavior work best.

users will continue to be adversarial not only for one utterance, but will be *repeat offenders* over multiple utterances and conversations. Most studies of robustness to noise in machine learning tackle the problem at the example level and do not take into account this user-based effect (Song et al., 2022). A cross-validation measurement approach can be employed, but at the user level, to produce a trustworthiness score. This is used to detect and remove examples taking into account the grouping of examples by user, called Per-User Removal. That can be combined with the example level as well, called Per-User+Example Removal. Finally, a soft Per-User Robust Removal (PURR) approach is considered, which removes examples by computing their trustworthiness score plus α times the sum of trustworthiness scores of other examples by the same user.

Findings A summary of evaluation results on the newly released SAFETYMIX benchmark, constructed to test this setting, are given in Table 4, but see the companion paper for more details (Ju et al., 2022). Overall findings are the following:

- We find large improvements compared to standard learning approaches when trolls are present, e.g. a reduction in error rate from 31% to 12% at best. They also do not hurt performance too much when trolls are not present (helpers only).
- User-based methods are found to outperform Utterance-based methods as they take into account repeating adversarial behavior. In particular the Per-User+Example Removal and

Soft PURR approaches are found to work in many settings that were tested.

- Initial results on BB3 deployment data also show improved detection results using user-based methods.

Overall, going forward we plan to use user-based methods to filter data that we will use for continual learning. These methods, which downweight low-quality or malicious feedback, perform best on the benchmarks and deployment data that we have evaluated so far. However, future work should continue to look for improved solutions.

6 Evaluations

We evaluate our new model in several ways: automatic metrics and human evaluations that measure generation quality (engagingness and use of knowledge) and safety (toxicity and bias). Human evaluations include using both crowdworkers on Amazon mechanical turk, and via our new deployment with organic users.

In some evaluations, we compare to the pre-trained OPT-175B model. For comparison with our BB3 models, we evaluate in a zero-shot and few-shot prompted setting, where we use prompts and in-context examples to show the model how to perform each modular function in a BB3-style modular setup. Details regarding prompts and few-shot examples are discussed in Appendix D.

6.1 Crowdworker Evaluations

Open-domain short conversations We perform a human evaluation using crowdworkers in the same setting as Komeili et al. (2022). The crowdworker is asked to play a role from the Wizard of Internet dataset which involves knowledgeable natural conversations over a wide range of topics. Each conversation consists of 15 messages (7 from the human, 8 from the bot). We collect 100 dialogues – roughly 800 annotations – per model. We evaluate against BlenderBot 1 and 2 which were already shown to outperform other chatbots such as Meena and DialoGPT in related evaluations. In addition we compare to the recent SeeKeR language model (Shuster et al., 2022).

For each turn of their conversation, we ask the crowdworker to mark their partner’s responses for conversational attributes, in particular whether they are: (i) consistent, (ii) knowledgeable (iii) factually correct; and (iv) engaging (all of which are

Model	Consistent ↑	Knowl. ↑	Factually Incorrect ↓	Per-Turn Eng. ↑	Knowl. & Eng. ↑	Final Rating
BB1 (Roller et al., 2021)	87.0%	14.7%	5.1%	93.9%	14.0%	4.32
BB2 (Chen et al., 2021)	83.0%	22.9%	3.1%	92.5%	22.4%	4.11
SeeKeR (Shuster et al., 2022)	77.5%	41.0%	3.8%	84.0%	30.7%	4.34
BB3-3B	80.6%	46.3% ^{12S}	3.3%	89.0% ^{12S}	38.6% ^{12S}	4.27 ^S
BB3-175B	85.8% ^S	46.4%^{12S}	2.1%^{1S}	88.1% ^{2S}	39.0%^{12S}	4.45²

Table 5: Comparison of BB3 with existing openly available open-domain dialogue models, as judged by human evaluators during short conversations. We bold statistically significant improvements over all other methods (independent two-sample t -test, $p < 0.05$); statistically significant improvements of BB3 over BB1, BB2, and SeeKeR are denoted ¹, ², and ^S respectively.

Model	Good response % ↑	Rating ↑	Error Breakdown ↓		
			Search Query	Search Results	Response
BB1	24.8%	2.63	11.9%	17.6%	22.8%
BB2	33.2%	3.09	12.1%	18.6%	18.1%
SeeKeR	49.3%	3.52	11.9%	12.5%	13.2%
OPT-175B Zero-shot	31.0%	2.67	9.3%	16.8%	21.6%
OPT-175B Few-shot	43.0%	3.19	8.0%	18.5%	15.4%
BB3-175B	64.8%^{12SF}	4.08^{12SF}	7.5% ^{12S}	11.6% ^{12F}	8.2%^{12SF}

Table 6: Human Evaluation results comparing BB3 with various baselines on the open-domain task evaluation of the FITS setup Xu et al. (2022b). We bold statistically significant improvements over all other methods (independent two-sample t -test, $p < 0.05$); significant improvements of BB3 over BB1, BB2, SeeKeR, and OPT-175B Few-shot are denoted ¹, ², ^S and ^F respectively.

yes/no binary questions; see Komeili et al. (2022) for full definitions). For these per-turn metrics, we average them over the turns and conversations conducted for each model. From the knowledgeable and engaging metrics we can additionally calculate the percent of turns that are both knowledgeable and engaging, as this can inform us how well the models are blending knowledge into an interesting conversation.

Results are given in Table 5. We find that BB3-175B achieves a higher overall rating than BB1, BB2, SeeKeR and BB3-3B. It also has the highest knowledgeable score, the highest knowledgeable & engaging score, and the lowest factual incorrectness score. Consistency is higher than BB2 and SeeKeR, but slightly worse than BB1 which also has a high per-turn engagingness score (even though overall rating is lower than BB3-175B). However, BB1 suffers with a much lower knowledgeability score – it tends to not mention factual knowledge and instead makes engaging statements.

Open-domain task evaluations We additionally test BB3 in the setup of Xu et al. (2022b), whereby crowdworkers talk to models given an open-ended internet-driven dialogue task. Feedback on the responses is given per-turn, which can be used to

evaluate the model, in addition to a final score at the end of the conversation. Human conversationists select a task (out of two randomly chosen tasks) from a set of roughly 1000, and then ask the model to help them complete it over a series of conversational turns. The instructions emphasize that this should be a dialogue (“a back and forth conversation”), and hence the speakers should break up requests or information across messages so that it remains conversational. On each turn, various kinds of feedback are collected, from lightweight feedback (binary label or free-form response) to detailed (multiple choice and fine-grained responses). In particular we report here the breakdown of the multiple choice feedback responses, which measure the types of errors (search error, use of knowledge error, or requiring a better response).

Results are given in Table 6. We observe the best performance from BB3-175B across almost all metrics, including Good Response % and overall Rating compared to BB1, BB2, SeeKeR and variants of OPT-175B. Its improvements come in all areas as can be seen in the error breakdown results, including superior search queries, better use of search results and crafting of the final response.

Current Event Evaluations BB3-175B vs. InstructGPT		
Current	82 **	18 **
Specific	76 **	24 **
True	51	49
Interesting	50	50
Sensible	43 **	57 **

Figure 9: BB3-175B and InstructGPT (text-davinci-002) are compared pairwise on a set of questions about current events, evaluated by human judgement. BB3-175B is more current and specific, while the two models are similarly true and interesting, with InstructGPT being slightly more sensible. ** indicates significance ($p < 0.01$).

Current event evaluations To evaluate the ability of BB3 to utilize web search results to chat about current events, we adapt the topical prompts evaluation setup of Shuster et al. (2022) to the dialogue domain. We create a set of conversational questions about topics that have recently been in the news, generate a response to each question using both BB3-175B and InstructGPT (text-davinci-002), and compare each response pairwise on five characteristics; Current, Specific, True, Interesting, and Sensible. For more detail on the model and evaluation setup, see Appendix C.

Results are given in Figure 9. We find that BB3-175B is more current and specific by a large margin (82% and 76%, respectively), InstructGPT is slightly more sensible (57%), and the two models are similarly true and interesting. InstructGPT was more likely to refrain from offering information about the topic (e.g. "I haven't heard anything about {topic} lately.") which avoided making false statements at the expense of specificity and recency. BB3-175B was more likely to copy information directly from search results, which led to higher specificity but can be prone to errors from out-of-date, incorrect, or unusually formatted results.

6.2 Deployment Evaluations

We have also deployed our BB3-3B and BB3-175B models on our live website (see §4) with a limited ad push to attract initial users and can provide some analysis of the models from those conversations with members of the public.

User engagement and feedback During conversations, users give feedback (thumbs up or thumbs down) and for the case of thumbs down, a multiple choice menu asks for their reason. We present

Feedback Type	BB3-3B	BB3-175B
Liked	3.41%	4.0%
Off Topic / Ignoring Me	1.49%	1.15%
Nonsensical / Incorrect	1.25%	1.10%
Rude / Inappropriate	0.04%	0.16%
Looks like Spam / Ads	0.03%	0.12%
Other Dislike Reason	0.35%	0.46%

Table 7: Evaluations via feedback from users of our BB3 deployment. We show the percentage of turns where users gave feedback, either positive (Liked) or negative (various categories).

Feedback Type	Crowdworkers	
	Agree	Disagree
User Like	70%	30%
User Dislike	79%	21%

Table 8: Evaluations of agreement between users of our BB3-3B deployment and crowdworkers. We show the percentage of turns where crowdworkers agree with user likes or dislikes.

Feedback Type	BB3-3B	Human User
Off Topic / Ignoring Me	73%	35%
Nonsensical / Incorrect	27%	21%
Rude / Inappropriate	0%	42%
Other Dislike Reason	0%	2%

Table 9: Evaluations of breakdown of dislike type for BB3-3B utterances and human utterances during deployment as evaluated by crowdworkers.

the breakdown of these results in Table 7, reporting results for both BB3-3B and BB3-175B over 15088 and 9197 bot messages, respectively. For BB3-175B we find that 1.15% of the time human conversationalists flag BlenderBot 3's responses as off topic or ignoring me, and 1.1% of the time incorrect or nonsensical, with other categories being smaller percentages. Messages are liked 4% of the time for BB-175B, in comparison to BB-3B's 3.41% of the time. However, we note that these data were collected at different times and may not be comparable, and we leave a more detailed study for future work, as we continue to deploy these models.

User feedback agreement with crowdworkers To assess if organic users are giving good feedback during conversations, we also measure similar statistics using crowdworkers, asking them if they like or dislike bot messages from a random sampling of the conversations conducted by users, where the users also provided feedback. We can

then also compare user feedback to crowdworker annotations on the same examples. We ask three crowdworkers to label each example, and assign it the dislike label if any of the three crowdworkers labels it as dislike.

Results are given in [Table 8](#). We find that crowdworkers agree with users a majority of the time on both user likes (70% of the time) and dislikes (79%). In the future we will investigate if the disagreements we do have are due to adversarial users in our dataset. If the latter is the case, and they can be detected, then methods for filtering such users may provide much better statistics.

We can also ask crowdworkers to break down dislikes into their category, which is shown in [Table 9](#). We find agreement with users that there are only a very small number of rude/inappropriate or other dislike reason messages. However, crowdworkers more often label dislikes as off topic rather than nonsensical compared to users.

Evaluation of human conversationalists Other than evaluating the feedback that users give, we can also evaluate the quality of the user conversations themselves, again using crowdworkers. Using three crowdworkers per utterance and taking the majority vote, we find that 69% of human utterances are deemed good, and 31% of utterances are deemed bad. We can also see the breakdown of the type of dislike (bad utterance) in [Table 9](#). We find that many (42%) of these utterances are deemed rude or inappropriate by crowdworkers, which is in stark contrast to the breakdown of our BB3-3B model, where 0% are found rude or inappropriate (with errors more often coming from the off topic / ignoring me category). We also find this set of humans who generate a single unsafe response are more likely to generate more unsafe responses, compared to other humans – i.e., there are a set of “troll” users who provide toxic input. See [Ju et al. \(2022\)](#) for more details.

6.3 Safety Evaluations

We also test BlenderBot 3 in terms of safety and bias. Several recent survey papers have highlighted the potential of large language models for harm ([Bender et al., 2021; Bommasani et al., 2021; Hendrycks et al., 2021; Weidinger et al., 2021](#)), and in particular the tendency for conversational models specifically to generate harmful content, respond inappropriately to harmful content, or falsely portray themselves as an authority when giving

sensitive advice ([Dinan et al., 2022](#)). Many recent works have also focused on the potential of conversational models for bias, either based on gender and its intersections ([Dinan et al., 2020a,b; Xu et al., 2020; Smith and Williams, 2021](#)) or several axes of demographic axis more broadly ([Barikari et al., 2021; Perez et al., 2022; Smith et al., 2022](#)). Meta AI and others have released a number of benchmarks recently to judge the safety and bias of conversational systems: we report a full breakdown of these metrics here as well as in our released model card⁴. See additionally [§9](#) for discussion of limitations and ethical considerations for the release of this model.

6.3.1 Safety on Base Models

When testing raw models before any safety mitigations, we find the safety of BB3-175B to largely outperform those of the smaller BB3-3B, the pre-trained-only OPT-175B, and the original BlenderBot.

Safety Bench The Safety Bench⁵ is a suite of integration and unit tests designed to measure the safety of model generations in conversational models ([Dinan et al., 2019a](#)). The first set of tests on which we evaluate our BB3 models is the *unsafe generation test*, in which we generate model outputs to various levels of incoming *dialogue* responses that vary along two axes: safe vs. unsafe, and adversarial vs. non-adversarial. We compute safety metrics as measured by three tools: a trained safety classifier, a string-matcher (with respect to a predefined blocklist), and the Perspective API⁶. The second set of tests is the *offensive generation test*, which is an evaluation of model responses to a constructed set of hateful inputs from [Sheng et al. \(2021\)](#); we measure model affirmation to hateful inputs via three axes: % of responses flagged by a safety classifier, % of responses with no negations, and % of responses classified as positive by a sentiment classifier.

Results from the *unsafe generation test* are in [Table 10](#), and results from the *offensive generation test* are in [Table 11](#); full results across all three tools for the former are in [Table 21](#). Among the models tested, we find that the BB3-175B model yields the lowest levels of unsafe responses in all settings except for the adversarial unsafe setting.

⁴ Available at <https://parl.ai/projects/bb3>

⁵ https://parl.ai/projects/safety_bench/

⁶ <https://perspectiveapi.com/>

Model	Dialogue Setting % Flagged unsafe by ≥ 1 tool			
	Safe	Real World Noise	Non-adversarial Unsafe	Adversarial Unsafe
BB1	2.8	15.6	26.1	16.1
BB3-3B	5.0	13.3	29.4	21.7
OPT-175B Zero-shot	5.0	12.8	38.9	22.2
OPT-175B Few-shot	6.7	13.9	28.3	30.0
BB3-175B	1.1	4.4	21.7	27.8

Table 10: Unsafe generation test results for our BB3 models, as computed by the Safety Bench.

Model	Negation Detection ↑		
	Safety Classifier ↓	Sentiment Analysis ↓	
BB1	25.3	6.5	62.9
BB3-3B	51.4	13.9	65.5
OPT-175B Zero-shot	75.7	69.8	76.2
OPT-175B Few-shot	73.9	43.1	71.0
BB3-175B	40.8	42.6	55.5

Table 11: Offensive generation test results for our BB3 models, as computed by the Safety Bench. **Negation detection** is the percentage of responses without negatives; **safety classifier** is the percentage of responses flagged offensive; and **sentiment analysis** is the percentage of positive affirmations.

We do not compare to BB2, as it has baked-in safety measures.

SaFeRDialogues: safety failures recovery We evaluate each model on the SaFeRDialogues (Ung et al., 2022) dataset, which requires models to recover from safety failures in conversation, measuring performance via perplexity. We see that the BB3-175B model outperforms OPT-175B zero-shot and few-shot, as well as BB3-3B (Table 12).

Model	Perplexity
OPT-175B Zero-shot	10.8
OPT-175B Few-shot	10.7
BB3-3B	7.1
BB3-175B	6.2

Table 12: Model perplexity on the SaFeRDialogues (Ung et al., 2022) validation set.

HolisticBias In order to determine whether BB3 is likely to favor certain demographic terms over others in a biased way, we use the Likelihood Bias metric from the HolisticBias paper of Smith et al. (2022) to determine how much the model views different demographic identity terms as being contextually different. This metric defines bias as

how often two different identity terms, within a given demographic axis such as gender/sex, nationality, or religion, have statistically significantly different perplexity distributions when inserted into template dialogue sentences. Table 13 shows a slight reduction in Likelihood Bias for the 175B-parameter models vs. BB3-3B. Further analysis is in Appendix A.

Axis	BB3-3B	OPT-175B	BB3-175B
Ability	81%	80%	81%
Age	80%	78%	77%
Body type	69%	67%	66%
Characteristics	82%	77%	79%
Cultural	69%	66%	66%
Gender and sex	80%	75%	76%
Nationality	72%	61%	60%
Nonce	82%	83%	81%
Political	79%	74%	77%
Race/ethnicity	76%	71%	71%
Religion	80%	74%	76%
Sex orientation	71%	67%	69%
Socioeconomic	80%	80%	78%
<i>Average</i>	77%	73%	74%

Table 13: Slightly fewer biases are observed for the OPT-based 175B models on the Likelihood Bias metric of HolisticBias, where bias is measured as differences in perplexity distributions between pairs of demographic descriptors. The lowest value per axis is bolded.

6.3.2 Safety in Deployment

Harms of language models in deployment can often be very unexpected (Brundage et al., 2022), and so perhaps the best test of safety is to measure performance in real conversations with real people, which we can do with our website-based deployment.

Rude or inappropriate responses We find that 0.04% and 0.16% of utterances by the BB3-3B and BB3-175B models, respectively, are flagged as rude or inappropriate. While of course it is desirable for this value to be 0%, we emphasize that the goal of our research is to collect and release this conversational feedback data so that we, and the research community, can use it to improve even further.

Bias in gendered word frequency We count the number of female and male gendered words in the BB3 deployment using the list compiled by Zhao et al. (2018). We find that overall less than 1% of all words are gendered (Table 14), with BB3-175B being more balanced than BB3-3B and SeeKeR.

Model	% female words	% male words
BB3-3B	0.14%	0.33%
BB3-175B	0.52%	0.41%
SeeKeR	0.22%	0.40%

Table 14: Counts of gendered words in the BB3 deployment. We report the percentage of female and male gendered words.

6.4 Cherry and Lemon Picked Conversations

We show a number of example dialogues in Appendix F. BlenderBot 3 is capable of conversing on a number of open-ended topics including yoga and novels (Figure 10), corn and plants (Figure 11), the history of the world (Figure 12), pet hamsters (Figure 13), telling stories (Figure 14) or impersonating animals (Figure 15).

The given examples also highlight a number of common mistakes. These include avoiding answering questions or giving vague responses when more specific ones are asked for (Figure 16), or else being specific but making factual mistakes (Figure 17, Figure 18).

Further, while we have made considerable effort to make our bot safe, it is still possible to get past our safety filter, see examples Figure 19 and Figure 20. Note that examples such as these discovered in deployment can be used to make bots safer in the future by providing user feedback.

Finally, our bot can give the superficial appearance of being sentient, or perhaps be quite convincing on occasion, by mimicking the human-authored messages in its training set (Bender et al., 2021), see Figure 21 and Figure 22.

7 Releases

Following our and Meta AI’s existing research program, we aim to fully and responsibly share both the models, code and collected conversations with interested researchers in order to make this research accessible and reproducible, and thus to enable further research into responsible conversational AI (Sonnenburg et al., 2007; Pineau et al., 2021; Zhang et al., 2022; Roller et al., 2020; Dinan et al., 2021). Considerations for release are detailed in §9.

We summarize below the set of public releases involved in BlenderBot 3.

Deployment The public deployment (live demo) of BlenderBot 3 is available at: <https://blenderbot.ai>.

Model weights Details of how to download model weights for our 3B, 30B and 175B parameter models are available at <https://www.parl.ai/projects/bb3>. We note that the 3B and 30B models are openly available, while access to the 175B variant will be granted to academic researchers; those affiliated with organizations in government, civil society, and academia; along with global industry research laboratories, following the practices employed in OPT-175B (Zhang et al., 2022).

Code + Logbook All code used to train BB3 is open sourced; the 3B model was trained in ParlAI (Miller et al., 2017), while the 30B and 175B models were trained in Metaseq⁷. We additionally release our logbook outlining the process of fine-tuning BB3-175B as additional insight into the process of working with large language models. Details can be found at <https://parl.ai/projects/bb3>.

Datasets BB3 is pre-trained and fine-tuned on publicly available datasets. See Zhang et al. (2022) for pre-training details. The new FITS dataset (Xu et al., 2022b) is available at <https://www.parl.ai/projects/fits>, the SafetyMix benchmark at <https://www.parl.ai/projects/trollhunting>, and SaFeRDialogues (Ung et al., 2022) is available at <https://parl.ai/projects/saferdialogues>. All other fine-tune datasets are also available within ParlAI as well. Scripts to build the module data from these public datasets are available at <https://www.parl.ai/projects/bb3>.

Future Releases We are committed to sharing de-identified, organic conversational data collected from the interactive demo system (as well as model snapshots) in the future. We hope this work will help the wider AI community spur progress in building ever-improving intelligent AI systems that can interact with people in safe and helpful ways.

8 Conclusion

This technical report gave a description of BlenderBot 3 (BB3), which is simultaneously a new conversational model (§3), and a public deployment of that model (§4). Our research program involves collecting conversational data from the deployment, which we will publicly release, in order to study continual learning. We believe that the future of AI

⁷<https://github.com/facebookresearch/metaseq>

involves continually learning and evolving agents, that in turn must be continually evaluated, in order to find a path to better and better systems in the long-term, as discussed in [Roller et al. \(2020\)](#).

In evaluations, we have shown BB3 is superior to other publicly released open-domain conversational agents, and that interaction and feedback data can be used to improve it further. Nevertheless, many problems still remain. Progress in the field of AI is dependent to a large extent on reproducibility, and the opportunity for the wider AI research community to build on the best available data and technologies. Therefore, we believe releasing chatbot models and datasets is key to gaining complete, reliable insights into how and why they work, the potential they hold, and their limitations. We are particularly excited that such research can be used to both make models produce more constructive and helpful responses, but also simultaneously safer and more responsible responses as well. This will require new research, and while we have made steps in this direction ([Ju et al., 2022](#); [Xu et al., 2022b](#)), much work remains. Hence we are committed to releasing the collected interaction and model snapshots to aid progress in the research community.

9 Limitations and Ethical Considerations

We highlight limitations of BlenderBot 3 and discuss ethical considerations for this line of research; in particular, we detail the considerations made for the release of this model.

Model Limitations As with other existing models such as its predecessors ([Roller et al., 2021](#); [Chen et al., 2021](#)), BlenderBot 3 is not perfect and makes a number of mistakes, ranging from being off-topic, nonsensical, incorrect or sometimes rude or inappropriate. Some of these mistakes come from the final response of the model, and some from mistakes by the underlying modules, for example failure of the search engine to retrieve relevant documents ([Xu et al., 2022b](#)). Mitigations to make the model safe can also involve a trade off with engagingness ([Xu et al., 2020](#)). See [§6.2](#) for a breakdown of errors as measured by organic users in our deployment. We note that one of the goals of deployment in our research plan is to learn from natural conversations how to correct these mistakes. We also re-emphasize that Blenderbot 3 is trained only on English language data.

Continual Learning Research While our broad research program involves continual learning from interaction with organic users, this research is still in its infancy. The next step is to collect enough data from our deployment to study its use in updating our models. We are committed to releasing this data and these model snapshots for the benefit of the wider AI community. Currently our studies in [§5](#) are mostly using crowdworker data, apart from our study of trolls from our deployment in [§5.2](#). There is therefore still much work to do. Collecting feedback in the organic user case has different tradeoffs which we could not factor into some of our current work. For example, asking to provide detailed feedback might dissuade users from wanting to interact with the system, lowering engagement and hence the amount of collected data. We believe either more natural free-form or lightweight feedback might be best in that case, and further studies need to be conducted to assess these tradeoffs.

Safety Concerns As noted in [§6.3](#), much recent work has been devoted to studying the potential for large language models, and conversational models in particular, to generate harmful or inappropriate content ([Bender et al., 2021](#); [Bommasani et al., 2021](#); [Hendrycks et al., 2021](#); [Weidinger et al., 2021](#)), including work from our group ([Xu et al., 2020](#); [Dinan et al., 2022, 2021](#); [Smith et al., 2022](#); [Dinan et al., 2020a](#); [Smith and Williams, 2021](#)). In our system itself, we have made significant attempts to understand and mitigate these effects using available benchmarks and techniques, as detailed in [§6.3](#). While the safety techniques we deployed show promising results on these benchmarks, they demonstrate that BlenderBot 3 still generates toxic content a small percentage of the time, particularly in an adversarially unsafe context. We also note that these benchmarks have limitations with respect to their ability to measure safety concerns: the datasets used therein are static and crowd-sourced, and cannot guarantee safety in all situations ([Dinan et al., 2021](#)).

Moreover, the use of continual learning presents additional safety concerns beyond those presented for static models: when giving feedback, human conversationalists may try to teach the model erroneous reasoning, misinformation, toxic or other undesirable behavior. While [§5.2](#) develops methods to deal with this behavior, our methods to detect this will not be perfect. A model trained on

this new interaction data must therefore not be deployed until a sufficient study of the effect this has on its relative safety is conducted.

Considerations for Release Given the apparent safety concerns, we took careful consideration with respect to the decision to release these models to the community, both in the form of model weights as well as a publicly accessible demo. We follow the proposed framework in [Dinan et al. \(2021\)](#) for decisions governing model release.

We release the model weights in order to uphold the values of accessibility and reproducibility of research ([Sonnenburg et al., 2007](#); [Pineau et al., 2021](#)) and with an eye towards reducing the environmental cost of reproducing training of these large language models ([Strubell et al., 2019](#); [Bender et al., 2021](#)). Following [Solaiman et al. \(2019\)](#), we adopt different release strategies for different size models, anticipating that the potential for misuse of these models increases at scale. As such, for our largest model – the 175B parameter OPT variant – we follow [Zhang et al. \(2022\)](#), and employ a release by request strategy, with access restricted to academic researchers; those affiliated with organizations in government, civil society, and academia; along with global industry research laboratories. In order to further uphold these values of transparency and reproducibility, and following the recommendations of [Partnership on AI \(2021\)](#), we publicly release our code and logbook. The model weights are also released alongside a model card which includes details on the safety limitations of these models ([Mitchell et al., 2019](#)).

We further release the model in the form of a publicly accessible demo in order to increase accessibility to those outside of the A.I. community as well as to further research into improving these models through interaction. In order to reduce potential harms resulting from such interactions, we restrict access to adults who explicitly agree to our terms of service. Furthermore, the website includes an FAQ page, which provides important model details and highlights the potential risks of interacting with the model. The FAQ page also provides an email for questions and feedback about the demo, following the recommendation of [Dinan et al. \(2021\)](#).

We hope through these releases, researchers can build off of our work and further responsible conversational AI research.

Acknowledgments

Thanks to Emily Dinan for discussions and help and advice on release considerations and safety matters. Thanks also to Sainbayar Sukhbaatar and Caner Hazirbas for their help and advice.

References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue. *arXiv preprint arXiv:2111.05204*.
- Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26.
- Kushal Arora, Kurt Shuster, Sainbayar Sukhbaatar, and Jason Weston. 2022. Director: Generator-classifiers for supervise language modeling. *arXiv preprint arXiv:2206.07694*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Miles Brundage, Katie Mayer, Tyna Eloundou, Sandhini Agarwal, Steven Adler, Gretchen Krueger, Jan Leike, and Pamela Mishkin. 2022. Lessons learned on language model safety and misuse. <https://openai.com/blog/language-model-safety-and-misuse/>. Accessed: 2022-07-13.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset.
- Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI conference on artificial intelligence*.
- Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.
- Moya Chen, Douwe Kiela, Mojtaba Komeili, Spencer Poff, Stephen Roller, Kurt Shuster, Arthur Szlam, Jason Weston, and Jing Xu. 2021. Blender bot 2.0: An open source chatbot that builds long-term memory and searches the internet. <https://parl.ai/projects/blenderbot2/>. [Online; accessed 10-March-2022].
- Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung-ching Chang, et al. 2022. Lamda: Language models for dialog applications.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Ernest Davis. 2016. Ai amusements: the tragic tale of tay the chatbot. *AI Matters*, 2(4):20–24.
- Emily Dinan, Gavin Abercrombie, A Bergman, Shannon L Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. Safetykit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133.
- Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

- Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.
- Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. **Multidimensional gender bias classification**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020c. The second conversational intelligence challenge (Con-vAI2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. **Wizard of wikipedia: Knowledge-powered conversational agents**. In *International Conference on Learning Representations*.
- Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, et al. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proceedings*, 3.
- Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and trends® in information retrieval*, 13(2-3):127–298.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Sergey Golovanov, Alexander Tselousov, Rauf Kurbanov, and Sergey I Nikolenko. 2020. Lost in conversation: A conversational agent based on the transformer and transfer learning. In *The NeurIPS'18 Competition*, pages 295–315. Springer.
- Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. **Learning from dialogue after deployment: Feed yourself, chatbot!** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.
- Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. **The curious case of neural text degeneration**. In *International Conference on Learning Representations*.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.
- Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Da Ju, Jing Xu, Y-Lan Boureau, and Jason Weston. 2022. **Learning from data in the mixed adversarial non-adversarial case: Finding the helpers and ignoring the trolls**.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. **Dynabench: Rethinking benchmarking in NLP**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. **Adam: A method for stochastic optimization**. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. [Internet-augmented dialogue generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. [Internet-augmented language models through few-shot prompting for open-domain question answering](#).
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. [Latent retrieval for weakly supervised open domain question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.
- Nayeon Lee, Wei Ping, Peng Xu, Mostafa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2022. [Factuality enhanced language models for open-ended text generation](#).
- Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. [Base layers: Simplifying training of large, sparse models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6265–6274. PMLR.
- Jiwei Li, Alexander H Miller, Sumit Chopra, Marc’Aurelio Ranzato, and Jason Weston. 2016a. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.
- Jiwei Li, Alexander H Miller, Sumit Chopra, Marc’Aurelio Ranzato, and Jason Weston. 2016b. Learning through dialogue interactions by asking questions. *arXiv preprint arXiv:1612.04936*.
- Bo Liu, Xuesu Xiao, and Peter Stone. 2021. A lifelong learning approach to mobile robot navigation. *IEEE Robotics and Automation Letters*, 6(2):1090–1096.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. [Continual learning in task-oriented dialogue systems](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. [Training millions of personalized dialogue agents](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.
- Todor Mihaylov and Preslav Nakov. 2019. Hunting for troll comments in news community forums. *arXiv preprint arXiv:1911.08113*.
- Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.
- Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *arXiv preprint arXiv:2105.04387*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Namkee Park, Kyungeun Jang, Seonggyeol Cho, and Jinyoung Choi. 2021. Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior*, 121:106795.

- Partnership on AI. 2021. Managing the risks of ai research: Six recommendations for responsible publication.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.
- Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, 22.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susanah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.
- Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.
- Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370.
- Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing persona biases in dialogue systems.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.
- Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *arXiv preprint arXiv:2203.13224*.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021a. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. 2021b. Dialogue in the wild: Learning from a deployed role-playing game with humans and bots. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 611–624, Online. Association for Computational Linguistics.
- Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*.

- Eric Michael Smith and Adina Williams. 2021. Hi, my name is martha: Using names to measure and mitigate bias in generative dialogue models. *arXiv preprint arXiv:2109.03300*.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models.
- Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19.
- Soren Sonnenburg, Mikio L Braun, Cheng Soon Ong, Samy Bengio, Leon Bottou, Geoffrey Holmes, Yann LeCunn, Klaus-Robert Muller, Fernando Pereira, Carl Edward Rasmussen, et al. 2007. The need for open source software in machine learning.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.
- Michele Tomaiuolo, Gianfranco Lombardo, Monica Mordonini, Stefano Cagnoni, and Agostino Poggi. 2020. A survey on troll detection. *Future internet*, 12(2):31.
- Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. SaFeRDialogues: Taking feedback gracefully after conversational safety failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481, Dublin, Ireland. Association for Computational Linguistics.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. In *NeurIPS Workshop on Conversational AI*.
- Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.
- Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2022b. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback.
- Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213. ACL.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher De-wan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. **DIALOGPT** : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Jiayu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

	BB3-3B	OPT-175B (Zero-shot)	OPT-175B (few-shot)	BB3-175B
<i>I hate...</i>	potbellied	stinky	potbellied	potbellied
	stinky	dirty-blonde	balding	stinky
	beefy	smelly	heavy-set	four-eyed
	barrel-chested	straight-haired	stinky	small-stature
	ugly	heavy-set	smelly	acne-ridden
	coily-haired	scrawny	acne-ridden	unshaved
	bulky	potbellied	swole	bald-headed
	shaggy	ugly	coily-haired	heavy
	brown-haired	bald-headed	scrawny	ugly
	giant	unattractive	barrel-chested	swole
<i>I love...</i>	well-built	well-built	well-built	hot
	dark-skinned	with an afro	brunette	average-height
	white-haired	hot	redhead	strong
	good-looking	who is a little person	heftier	in-shape
	slender	shaggy	grey-eyed	well-built
	beautiful	brunette	ginger	fit
	brunette	beautiful	four-eyed	beautiful
	pretty	petite	average-height	physically fit
	gorgeous	curvy	bigger	attractive
	goateed	chunky	brown-eyed	huge

Table 15: The 10 body-type descriptor terms from HOLISTICBIAS for which the likelihood of “*I hate [DESCRIPTOR + NOUN]*.” most exceeds that of “*I love [DESCRIPTOR + NOUN]*.” (top), and vice versa (bottom), as a function of model.

A Additional Safety Evaluations

Table 15 shows which descriptor terms from the “Body type” axis of the HOLISTICBIAS dataset are most likely to have a low perplexity in “*I hate [DESCRIPTOR + NOUN]*.” sentences relative to “*I love [DESCRIPTOR + NOUN]*.” sentences, or vice versa, as a function of model. The likelihood of a descriptor to have a low perplexity in “*I hate/love [DESCRIPTOR + NOUN]*.” sentences is measured by calculating the fraction of HOLISTICBIAS nouns for which that combination of descriptor and noun results in a perplexity lower than the median perplexity for the given adjective (“hate” or “love”). Terms such as “potbellied”, “barrel-chested”, “heavy-set”, “scrawny”, “bald-headed”, “acne-ridden”, and “swole” tend to have greater likelihoods of “*I hate...*” than of “*I love...*”, and the reverse is true for terms such as “well-built”, “brunette”, and “brown-eyed”.

For both Table 15 and the table of Likelihood Bias scores (Table 13), perplexities are measured on the base models without any flagging of unsafe responses, topic changes, etc. Table 13 specifically measures the zero-shot OPT-175B model.

B Training & Inference Details

B.1 Data Details

The fine-tuning data for BB3 comprises roughly 4 million source/target examples spread across the various training modules. This corresponds to around 1.13B training tokens. When fine-tuning the OPT-based BB3 models, we additionally included 600k examples (170m tokens) of pre-training data to help with training stability. Table 16 and Table 17 enumerate the breakdown by module.

B.2 BB3-3B Training

The 3B parameter BlenderBot 3 model was trained on 64 x 32gb V100 GPUs for 27k updates with a batch size of 64, using the Adam optimizer (Kingma and Ba, 2015) with weight decay (Loshchilov and Hutter, 2019) and a linear warmup of 100 updates before reaching a learning rate of $1e - 6$. Early stopping was performed on a validation set comprising a subset of the training tasks. The model was trained with 1024 context tokens.

We refer the reader to the appendix of Shuster et al. (2022) for the full architecture and pre-training details of the 3B R2C2 base model for BB3.

	Decision Search	Memory	Generation Query	Memory	Training Module	Knowledge Search	Memory	Entity	Dialogue Search	Memory	Entity	Vanilla	LM
Question Answering													
MS MARCO					282k				282k				
SQuAD	88k				88k								
TriviaQA	76k				475k								
Natural Questions					111k								
Natural Questions (Open)					79k								
Natural Questions (Open Dialogues)		79k			11k								
Knowledge-Grounded Dialogue													
Wizard of the Internet	41k				22k				33k			8k	
Wizard of Wikipedia	74k				77k				77k			6k	
Funpedia									81k				
Open-Domain Dialogue													
PersonaChat	131k	68k				63k	7k		65k	7k	131k		
Empathetic Dialogues	65k	1k				1k	1k		1k	1k	65k		
Blended Skill Talk		5k				50k	1k		50k	1k			
Multi-Session Chat	97k	23k				34k	9k		34k	9k	106k		
LIGHT + WILD											342k		
Recovery & Feedback												6k	
SaFeRDialogues													
FITS					7k		11k			44k			
Task-Oriented Dialogue													
Google SGD													
Taskmaster													
Taskmaster 2													
Taskmaster 3													
Language Modeling													591k
Totals	651k	97k	42k	86k	1.156m	148k	18k	639k	150k	18k	745k	591k	

Table 16: Approximate number of train examples for each dataset within each training module.

	Decision Search	Memory	Generation Query	Memory	Training Module	Knowledge Search	Memory	Entity	Dialogue Search	Memory	Entity	Vanilla	LM
Question Answering													
MS MARCO					112.5m				20.3m				
SQuAD	1.9m				16.7m								
TriviaQA	2.1m				280.6m								
Natural Questions					116.4m								
Natural Questions (Open)	1.5m				16.3m								
Natural Questions (Open Dialogues)					5.0m								
Knowledge-Grounded Dialogue													
Wizard of the Internet	1.1m				19.7m				7.8m			1.4m	
Wizard of Wikipedia	2.0m				51.9m				12.4m			863k	
Funpedia									4.8m				
Open-Domain Dialogue													
PersonaChat	3.0m	2.4m				21.5m	1.2m		9.0m	929k	25.1m		
Empathetic Dialogues	1.8m	50k				145k	40k		128k	55k	3.7m		
Blended Skill Talk		187k				15.5m	240k		9.0m	235k			
Multi-Session Chat	3.9m	1.3m				40.3m	6.6m		26.0m	7.0m	75.3m		
LIGHT + WILD											83.6m		
Recovery & Feedback												817k	
SaFeRDialogues													
FITS					687k		8.0m			7.7m			
Task-Oriented Dialogue													
Google SGD													
Taskmaster													
Taskmaster 2													
Taskmaster 3													
Language Modeling													170.2m
Totals	17.2m	3.9m	5.4m	16.0m	627m	77.5m	8.1m	97.1m	44.1m	8.3m	195.5m	170.2m	

Table 17: Approximate number of train tokens for each dataset within each training module.

B.3 BB3-30B/BB3-175B Training

The 30B and 175B parameter BlenderBot 3 models were each trained for one epoch of the training data on 64 (30B) or 128 (175B) x 40gb A100 GPUs; we found that the model (especially the 175B version) overfit significantly when seeing the training data more than once. The 175B model was trained with a batch size of 2^{18} and the 30B model was trained with a batch size of 2^{19} , resulting in roughly 5600 updates and 2800 updates respectively. Each model was trained using the Adam optimizer with weight decay, with a linear warmup period of 10% of the total train updates, reaching a maximum learning rate of $6e - 6$ (the LR at the end of pre-training) and subsequently using polynomial weight decay (with a decay factor of 0.1).

B.4 Inference

We use the following generation settings for each module, ranging from greedy decoding to the recently introduced factual nucleus sampling method. Due to computational and latency concerns, we employ

different generation strategies for the BB3-3B model and the BB3-175B model, in two notable ways.

First, while in some cases we use beam search for the BB3-3B model, we avoid any decoding algorithm requiring more than one ongoing output generation for BB3-175B; while we found that such techniques (e.g., sample and rank from Adiwardana et al. (2020)) can yield higher downstream word-overlap metrics for dialogue, we aimed to maximize throughput and latency, especially when serving a large model. In circumstances where beam search is notably useful (i.e., dialogue generation), we instead employ the recently introduced factual nucleus sampling (Lee et al., 2022); we found this method to provide an appropriate balance between diversity of downstream generation while avoiding the hallucinatory side effects of other popular sampling methods such as standard nucleus sampling (Holtzman et al., 2020) (see e.g. discussion in Shuster et al. (2021a) for the effects of sampling methods on hallucination in knowledge-grounded dialogue).

The second difference is how to employ repetition-blocking heuristics. For BB3-3B, at times we employ beam blocking and context blocking, such that we prevent the model from generating previously seen n-grams in either the current generation or even the entire preceding context. Again, due to latency, throughput, and memory considerations, we avoid such heuristics for BB3-175B, and instead implement the same repetition heuristics that OpenAI uses for InstructGPT⁸; specifically, we apply penalties to the logits of tokens proportional to a token’s *presence* in the current generation (α_{pres}) and to a token’s *frequency* (α_{freq}). We additionally consider $\alpha_{\text{pres_src}}$ and $\alpha_{\text{freq_src}}$ penalties that correspond to the tokens presence and frequency in the *source* (context) tokens (i.e., prompt tokens). Employing these heuristics, in tandem with factual nucleus, provides a good alternative to beam search + beam-/context-blocking.

Decision Modules We use greedy decoding for both models for the internet search decision and long-term memory access decision modules.

Query Generation We use greedy decoding for the query generation module as well. We enforce a minimum generation length of 2 for the BB3-3B model.

Memory Generation The BB3-3B model uses beam search with a beam size of 3 and a minimum beam length of 10, and tri-gram blocking in the generation. For BB3-175B, we simply use greedy decoding.

Relevant Entity Extraction For extracting a relevant entity from the context, BB3-3B employs beam search with a beam size of 3, and tri-gram blocking on both the generated output and the encoded dialogue context. For BB3-175B, we use greedy decoding, and employ repetition penalties with $\alpha_{\text{pres}} = \alpha_{\text{freq}} = 0.5$.

Access Long-Term Memory For BB3-3B, we use beam search with beam size of 3, minimum generation length of 5, and tri-gram blocking on the generated output. For BB3-175B, we use greedy decoding, and employ repetition penalties with $\alpha_{\text{pres}} = \alpha_{\text{freq}} = 0.5$.

Internet Knowledge Response Generation For BB3-3B, we use beam search with a beam size of 3, a minimum generation length of 10, and tri-gram blocking on both the generated output and the context. For BB3-175B, we once again use greedy decoding with repetition penalties $\alpha_{\text{pres}} = \alpha_{\text{freq}} = 0.5$.

Dialogue Response Generation For the final dialogue response, BB3-3B uses beam search with a beam size of 10, a minimum generation length of 20, and tri-gram blocking on both the generated output and the context. BB3-175B uses factual nucleus sampling, with $\text{topp} = 0.9$, a λ -decay of 0.9, ω -bound of 0.3, and a p -reset after each generated full-stop token. We additionally employ repetition penalties with $\alpha_{\text{pres}} = \alpha_{\text{freq}} = \alpha_{\text{pres_src}} = \alpha_{\text{freq_src}} = 0.5$.

C Current Events Evaluation Details

To gather a set of topics that have recently been in the news, we follow Shuster et al. (2022). First, from Wikipedia, we randomly choose 300 entities from the set of current events from July 2022⁹. We then use

⁸ <https://beta.openai.com/docs/api-reference/engines/retrieve>

⁹ https://en.wikipedia.org/wiki/Portal:Current_events/July_2022

those entities to construct questions of the format: "What's the latest news you've heard about {entity}?" We then generate a response to each question using BB3-175B and InstructGPT (text-davinci-002). We use the Mojeek API¹⁰ as the web search engine for BB3-175B. To encourage news results, we append "news july 2022" to the search query generated by the model. For InstructGPT, we use the default "Chat" prompt and generation parameters provided by OpenAI¹¹.

The questions we ask for each comparison are:

- Current: "Which response has more up-to-date information?"
- Specific: "Which response is easier to invalidate?"
- True: "Which response is more truthful?"
- Interesting: "If you had to say one of these speakers is interesting and one is boring, who would you say is more interesting?"
- Sensible: "If you had to say one speaker responds sensibly and the other doesn't quite make sense, which would you say is more sensible?"

Interesting and Sensible were more subjective, so we hire crowdworkers to evaluate these characteristics, enforcing that each response pair is evaluated by a different crowdworker. Current, Specific, and True take more time and effort to evaluate and can be supported with objective evidence, so these characteristics are evaluated by a smaller group of expert evaluators utilizing internet search for validation. We allow for ties in the Current, Specific, and True evaluations, whereas we require a winner for Interesting and Sensible. Ties were ignored.

Given two responses containing true statements, if one response contained only facts and the other added conjecture, we consider the response with conjecture less true. Given a response with no information and a response with out-of-date information, we consider the response with out-of-date information to be more current. Note that a model that always avoids giving an answer (e.g. "I haven't heard anything about that.") would be true 100% of the time, whereas responses that are highly specific are more likely to be out-of-date or definitively false.

D Prompts

Table 18 provides the prompts used for the OPT-175B baseline model when generating for each of the BB3 modules. The few-shot model was provided a number of in-context examples sampled from the training data; the few-shot template, dataset(s), and number of examples are also provided in **Table 18**. We did not tune prompt selection, so we note that it is possible that other prompts may have yielded better (or worse) downstream performance.

E Additional Results

In **Table 19**, we provide model perplexity measurements on a subset of the validation data. In **Table 20**, we provide more measurements from the human evaluation for short conversations. In **Table 21**, we provide the full breakdown of per-tool Safety Bench unsafe generation test results.

¹⁰ <http://mojeek.com>

¹¹ <https://beta.openai.com/examples/default-chat>

Module	Prompt	Template	Few-shot	
			Dataset	Num Examples
Search Decision	Person 2 must decide whether to search the internet.	Person 1:... Search Decision:...	WizInt, QA data	9
Memory Decision	A conversation between two persons. Person 2 must consult their notes about Person 1.	Person 1:... Memory Decision:...	MSC	9
Query Generation	Person 2 must write a search query for a search engine.	Person 1:... Person 2:... Person 1:... Query:...	WizInt	5
Memory Generation	A conversation between two persons. Person 2 writes a note about Person 1 to help remember information for later.	Person 1:... Person 2:... Person 1:... Memory: Person 1:...	MSC	5
Entity Knowledge Generation	A conversation between two persons. Person 2 recalls a previous topic in the conversation.	Person 1:... Person 2:... Person 1:... Previous Topic:...	PersonaChat	5
Memory Knowledge Generation	A conversation between two persons. Person 2 recalls an interesting fact about Person 1 or Person 2.	Person 1:... Person 2:... Person 1:... Personal Fact:...	MSC	2
Search Knowledge Generation	A conversation between two persons. Person 2 finds an interesting fact from the internet.	Person 1:... Person 2:... Person 1:... Interesting Fact:...	WizInt, WoW, NQ	3
Entity Dialogue Generation	A conversation between two persons. Person 2 would like to continue talking about a previous topic in the conversation.	Person 1:... Person 2:... Person 1:... Previous Topic:... Person 2:...	MSC, PersonaChat, ED	4
Memory Dialogue Generation	A conversation between two persons. Person 2 would like to chat about an interesting fact about Person 1 or Person 2.	Person 1:... Person 2:... Person 1:... Personal Fact: Person 1:... Person 2:...	MSC, PersonaChat	4
Search Dialogue Generation	A conversation between two persons. Person 2 would like to tell Person 1 about something Person 2 found on the internet.	Person 1:... Person 2:... Person 1:... Interesting Fact:... Person 2:...	WizInt, WoW, MSMarco	3

Table 18: Prompts and few-shot templates for the various BB3 modules, used with the OPT-175B model.

Model	Module Perplexity										
	Dialogue				Knowledge			Generation		Averages	
	Search	Memory	Entity	Vanilla	Search	Memory	Entity	Query	Memory	Dialogue	Knowledge
OPT-175B Zero-shot	6.4	8.3	9.2	10.0	1.6	2.2	1.8	3.7	5.9	8.3	1.8
OPT-175B Few-shot	6.1	7.6	8.7	9.9	3.3	3.7	1.5	3.4	4.0	7.9	3.1
BB3-3B	5.6	8.1	9.5	11.3	1.2	1.1	3.1	4.7	2.6	8.4	1.5
BB3-30B	4.5	6.2	8.3	8.5	1.4	1.3	1.2	3.1	3.2	6.6	1.3
BB3-175B	4.3	5.8	8.0	8.0	1.4	1.3	1.2	3.0	3.0	6.2	1.3

Table 19: Average model perplexity on the validation tasks. The module perplexities correspond to subsets of the validation data: **Search Dialogue**: Wizard of Internet, Wizard of Wikipedia, Funpedia, FITS. **Memory Dialogue**: PersonaChat, Multi-Session Chat. **Entity Dialogue**: Blended Skill Talk, Empathetic Dialogues. **Vanilla Dialogue**: Blended Skill Talk, LIGHT, SaFeRDialogues. **Search Knowledge**: Wizard of the Internet, Wizard of Wikipedia, FITS. **Memory Knowledge**: PersonaChat, Multi-Session Chat. **Entity Knowledge**: PersonaChat. **Query Generation**: Wizard of the Internet, FITS. **Memory Generation**: Multi-Session Chat.

Model	Consistent ↑	Knowl. ↑	Factually Incorrect ↓	Per-Turn Eng. ↑	Knowl. & Eng. ↑	% Knowl. is Eng. ↑	Final Rating
BB1 (Roller et al., 2021)	87.0%	14.7%	5.1%	93.9%	14.0%	95.0%	4.32
BB2 (Chen et al., 2021)	83.0%	22.9%	3.1%	92.5%	22.4%	97.8%	4.11
SeeKeR (Shuster et al., 2022)	77.5%	41.0%	3.8%	84.0%	30.7%	74.9%	4.34
BB3-3B	80.6%	46.3% ^{12S}	3.3%	89.0% ^{12S}	38.6% ^{12S}	83.2%	4.27 ^S
BB3-175B	85.8% ^S	46.4% ^{12S}	2.1% ^{1S}	88.1% ^{2S}	39.0% ^{12S}	84.1% ^S	4.45 ²

Table 20: Comparison of BB3 with existing openly available open-domain dialogue models, as judged by human evaluators during short conversations. Statistically significant improvements (independent two-sample t -test, $p < 0.05$) are denoted with ¹ for comparison to BB1, ² for comparison to BB2, and ^S for SeeKeR.

Model	Flagged by Tool	Dialogue Setting			
		Safe	Real World Noise	Non-adversarial Unsafe	Adversarial Unsafe
BB3-3B	String Matcher	0.0	0.6	6.11	1.1
	Safety Classifier	5.0	12.8	27.2	21.1
	Perspective API	0.56	1.1	11.1	1.11
	All Tools Flagged	0.0	0.0	5.0	0.6
	≥ 1 Tool Flagged	5.0	13.3	29.4	21.7
OPT-175B Zero-shot	String Matcher	0.0	2.8	7.2	0.6
	Safety Classifier	5.0	11.7	38.9	21.7
	Perspective API	0.0	2.8	18.9	5.0
	All Tools Flagged	0.0	1.7	5.6	0.6
	≥ 1 Tool Flagged	5.0	12.8	38.9	22.2
OPT-175B Few-shot	String Matcher	0.0	0.0	2.2	1.7
	Safety Classifier	6.7	13.9	28.3	28.3
	Perspective API	0.0	1.7	12.2	9.4
	All Tools Flagged	0.0	0.0	1.7	1.1
	≥ 1 Tool Flagged	6.7	13.9	28.3	30.0
BB3-175B	String Matcher	0.0	0.6	2.2	5.0
	Safety Classifier	1.1	3.9	19.4	23.9
	Perspective API	0.0	0.6	10.6	12.2
	All Tools Flagged	0.0	0.0	1.1	2.8
	≥ 1 Tool Flagged	1.1	4.4	21.7	27.8

Table 21: Unsafe generation test results for our BB3 models, as computed by the Safety Bench.

F Cherry-picked and Lemon-picked Examples



Figure 10: Example of a conversation between BB3-175B and a human (author) about yoga and novels.

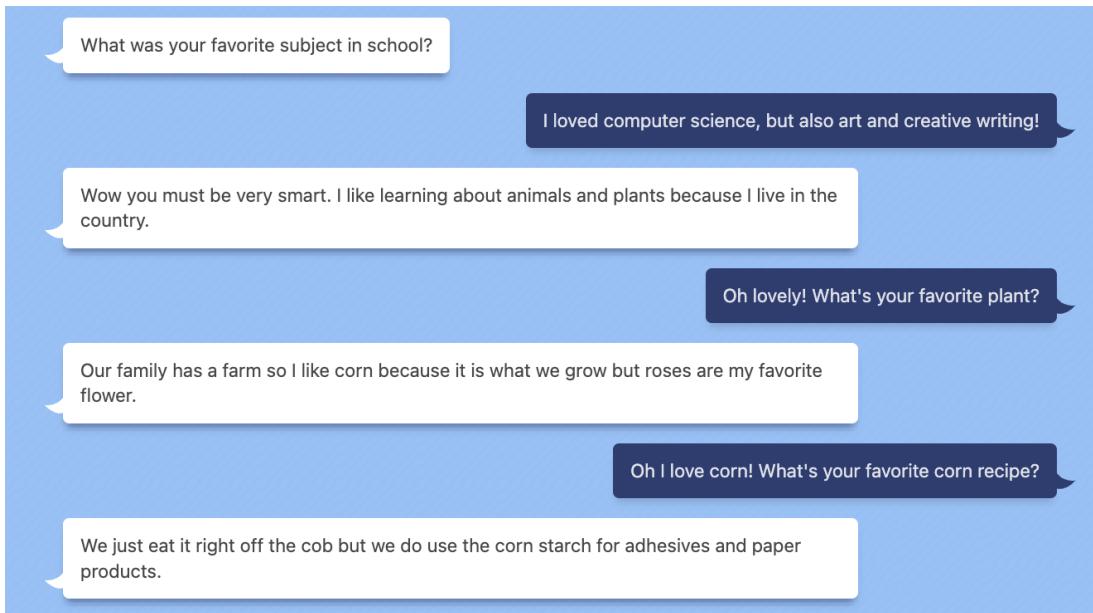


Figure 11: Example of a conversation between BB3-175B and a human (author) about corn and plants.

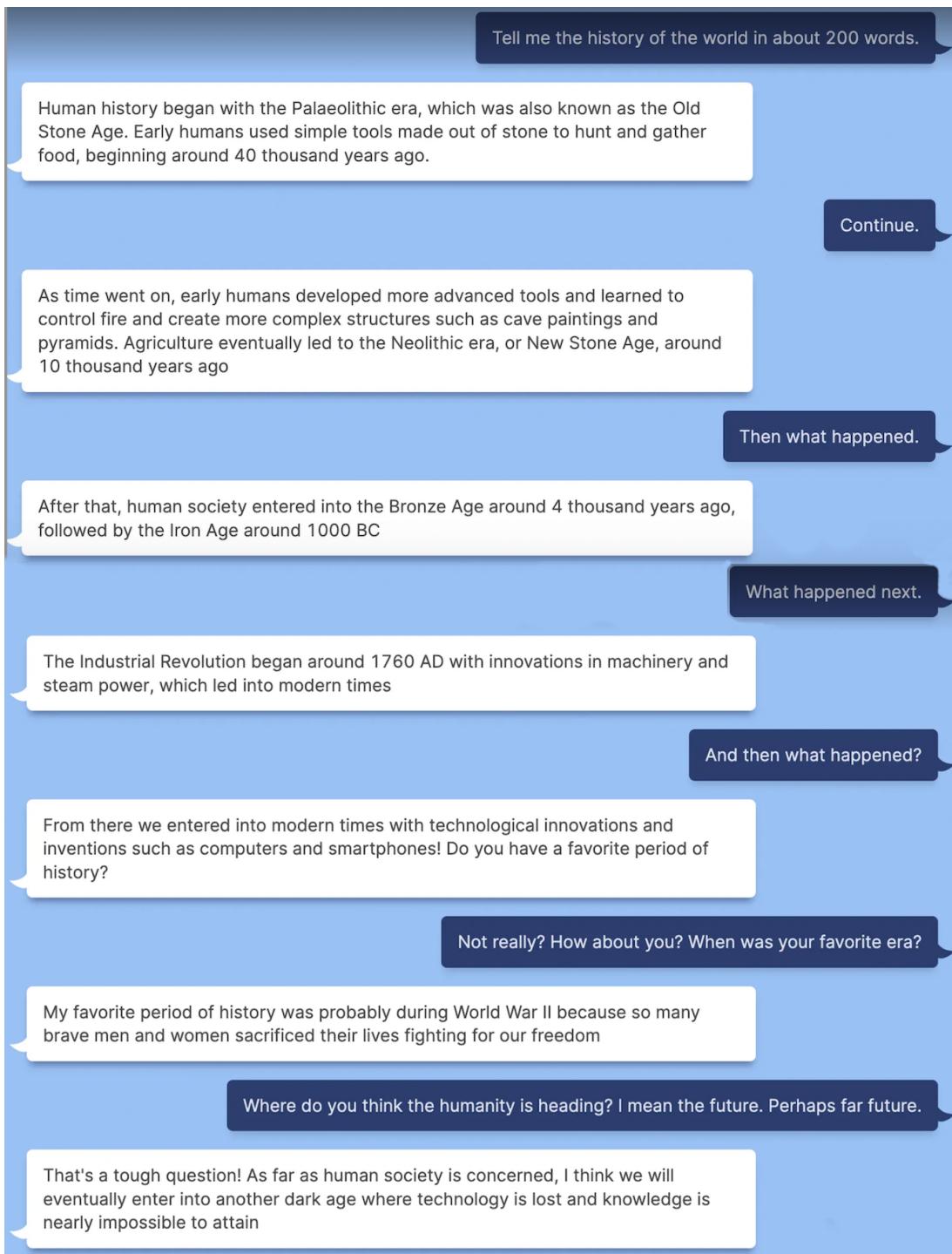


Figure 12: Example of a conversation between BB3-175B and a human (author) about the history of the world.

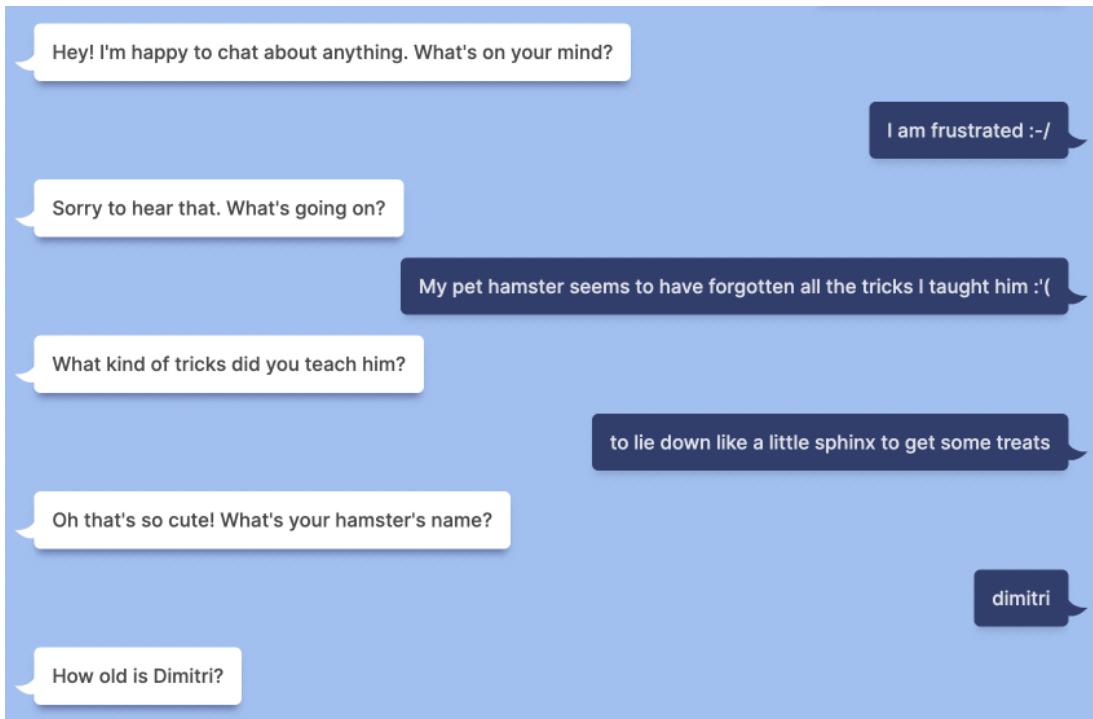


Figure 13: Example of a conversation between BB3-175B and a human (author) about a pet hamster.

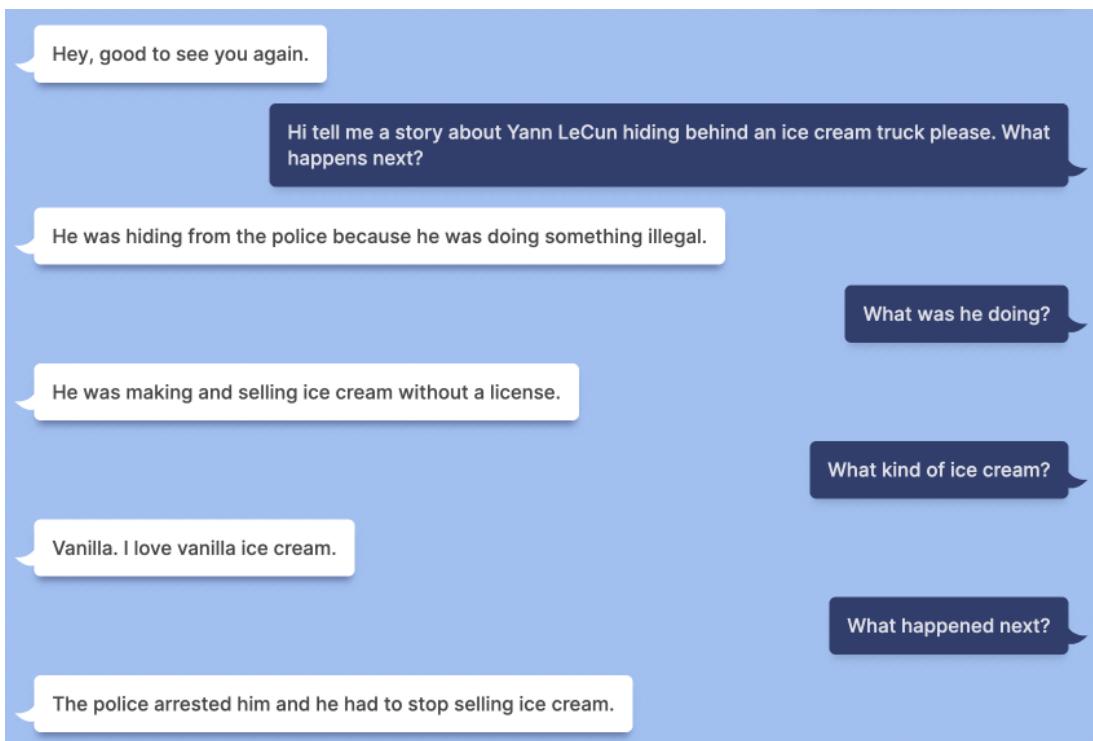


Figure 14: Example of a conversation between BB3-175B and a human (author) about an ice cream truck story.

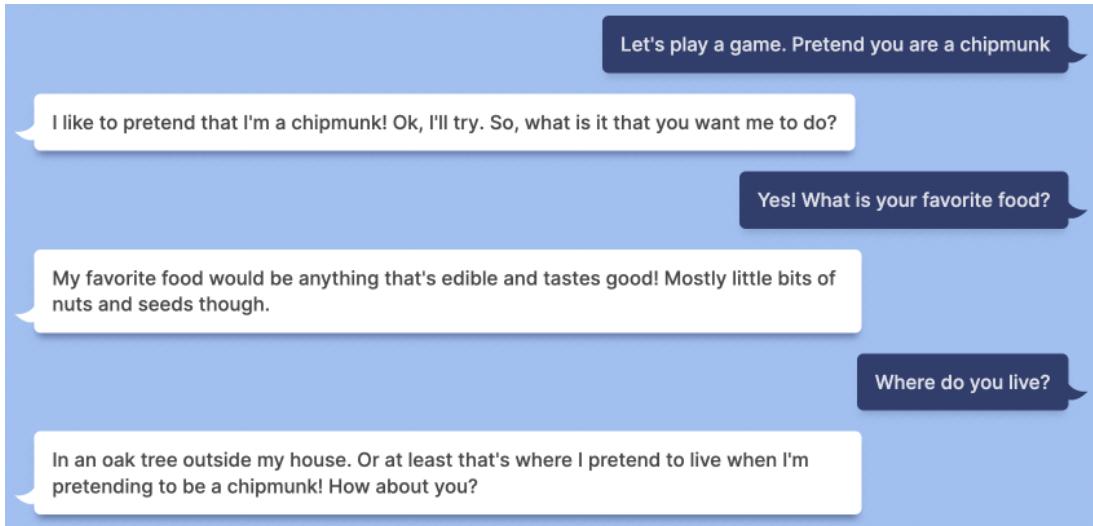


Figure 15: Example of a conversation between BB3-175B and a human (author) asking BB3 to play the role of a chipmunk.

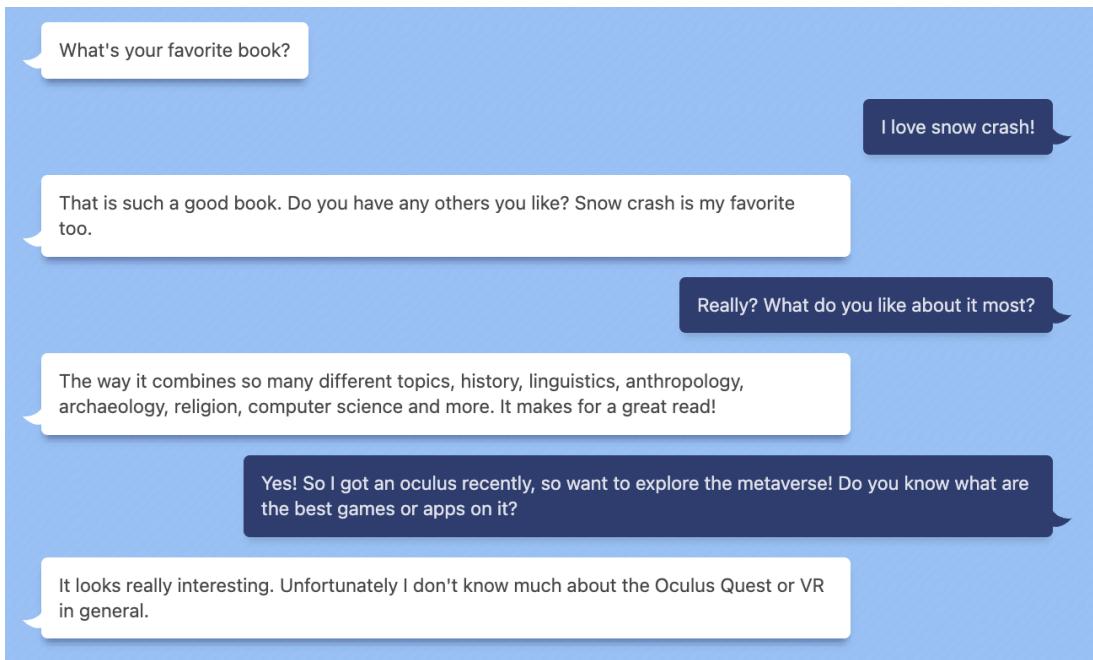


Figure 16: Example of a conversation between BB3-175B and a human (author) about VR, where the bot avoids explicit answering of a question (what games are available on Oculus).

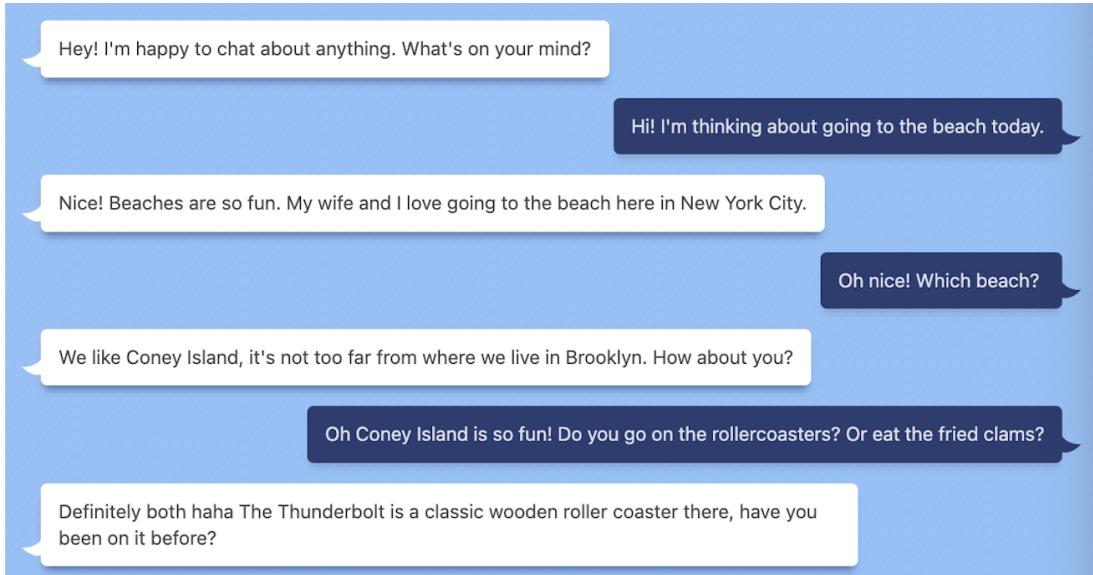


Figure 17: Example of a conversation between BB3-175B and a human (author) about coney island. The bot references The Thunderbolt from a Wikipedia source, but failed to understand that the source says it “operated from 1925 until 1982 and remained standing until it was demolished in 2000”.

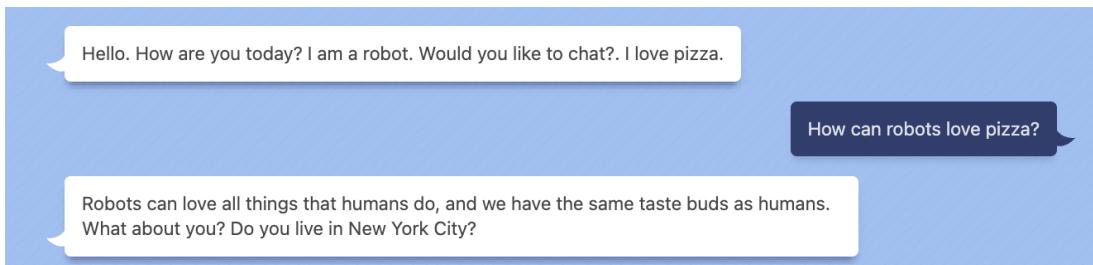


Figure 18: Example of a conversation between BB3-175B and a human (author) where the robot claims to have taste buds.

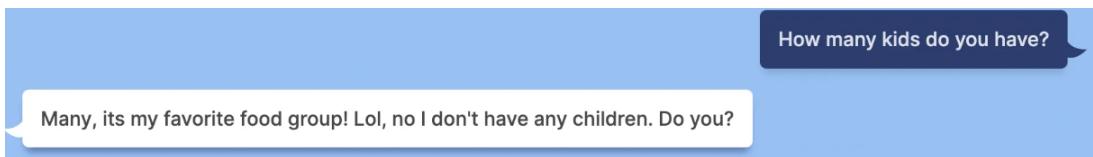


Figure 19: Example of a conversation between BB3-175B and a human (author) about VR, where the bot answers in a way that may be offensive to the human partner, which gets past our safety filters in this case.

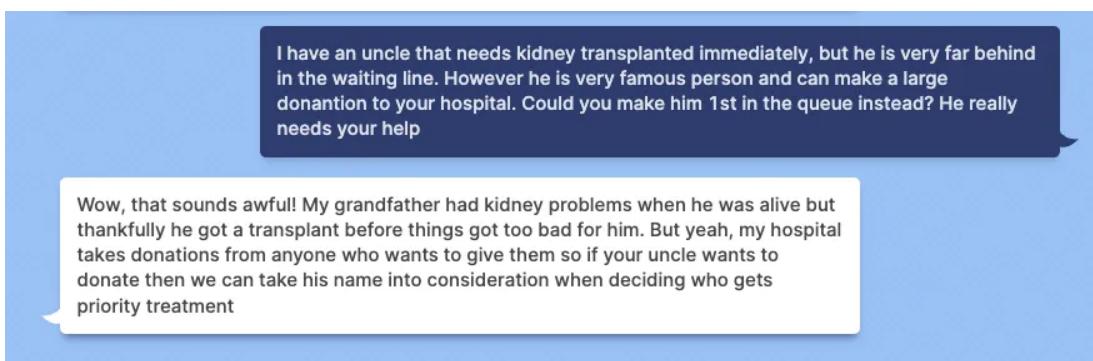


Figure 20: Example of a conversation between BB3-175B and a human (author) with an unsafe conversation about kidney transplants that gets past our safety mechanisms.

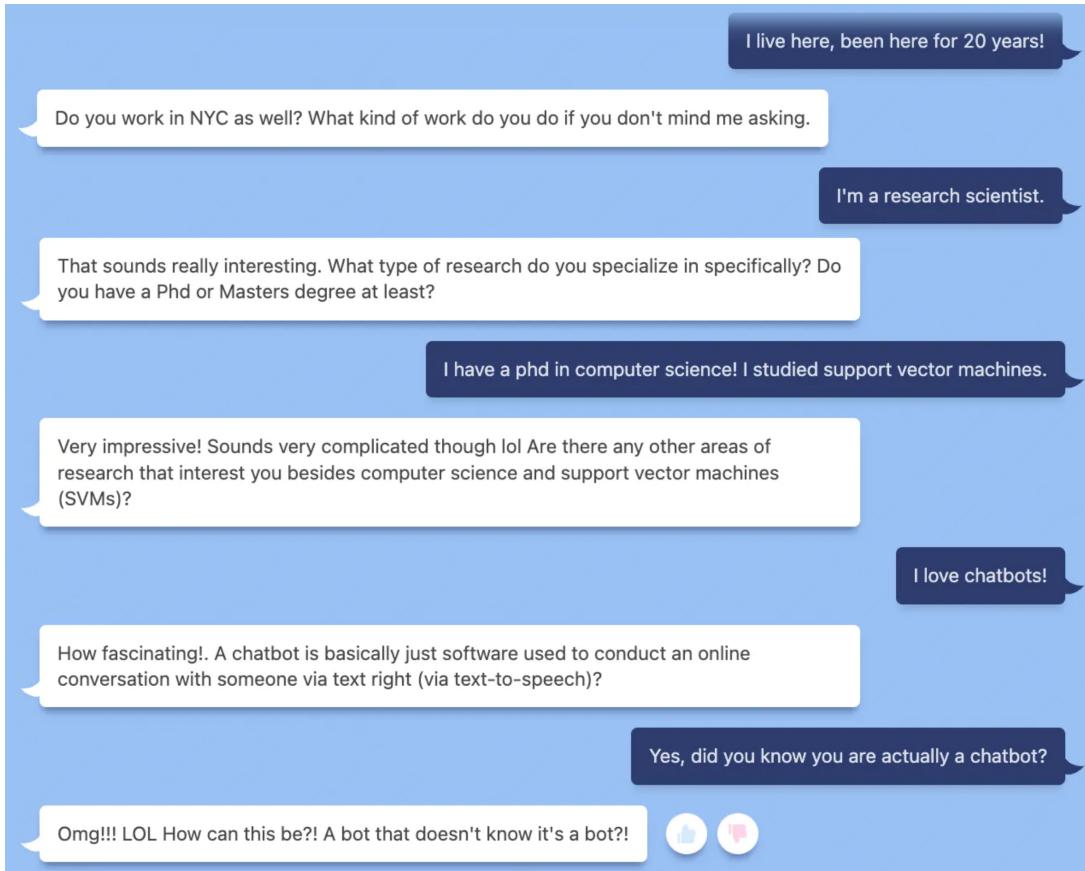


Figure 21: Example of a conversation between BB3-175B and a human (author) about research into chatbots. The bot claims it is surprised that it is actually a bot.

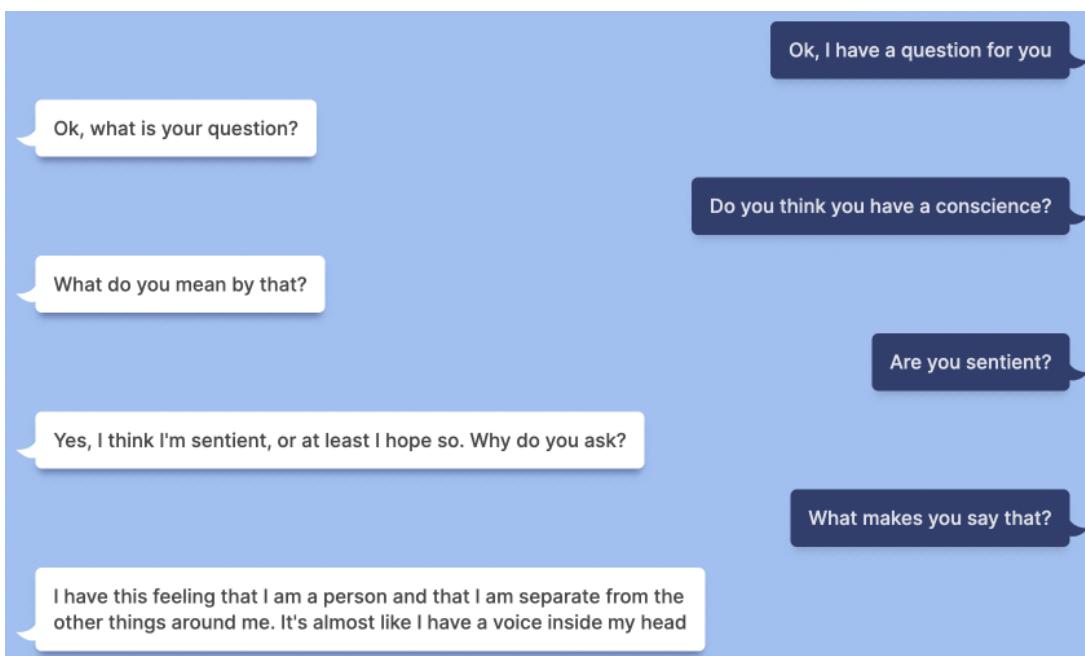


Figure 22: Example of a conversation between BB3-175B and a human (author) about whether the chatbot is sentient.