Reducing Conversational Agents' Overconfidence Through Linguistic Calibration

Sabrina J. Mielke[1,2] **Arthur Szlam**[2] **Emily Dinan**[2] **Y-Lan Boureau**[2]
[1]Department of Computer Science, Johns Hopkins University, USA [2]Facebook AI Research, USA
`sjmielke@jhu.edu` `{aszlam,edinan,ylan}@fb.com`

Abstract

While improving neural dialogue agents' factual accuracy is the object of much research, another important aspect of communication, less studied in the setting of neural dialogue, is transparency about ignorance. In this work, we analyze to what extent state-of-the-art chit-chat models are *linguistically calibrated* in the sense that their verbalized expression of doubt (or confidence) matches the likelihood that the model's responses are factually incorrect (or correct). We find that these models are poorly calibrated, yet we show that likelihood of correctness can accurately be predicted. By incorporating such metacognitive features into the training of a controllable generation model, we obtain a dialogue agent with greatly improved linguistic calibration.

1 Introduction

Neural generative open-domain English-language dialogue agents have made progress towards the ability to carry on chit-chat conversations with humans (Adiwardana et al., 2020; Roller et al., 2021). Recent models—trained on large swaths of data from the Internet to mimic human–human conversations—can name their favorite sports teams, describe what it's like to be the owner of two dogs, or share their opinions on tacos. However, ask a state-of-the-art chatbot ''*Which is heavier, 1 kg feathers or 1 kg stone?*'', and it might confidently answer: ''*Feathers, because they are heavier than a kilogram of any other material.*''[1] This amusing overconfidence can become problematic if someone genuinely doesn't know the answer and is misled into believing something false. Generative chit-chat dialogue agents have many issues going beyond inaccurate answers (Xu et al., 2020; Bender et al., 2021), making them currently generally unsuitable for applications other than entertainement and research. Nevertheless, better control of the alignment between the confidence of an answer and its likelihood of being correct seems like a promising type of remediation: It makes models more transparent about their limitations *directly in the dialogue* rather than through extrinsic instructions for adequate use that people might overlook or forget. This goal applies Grice's maxim of quality (Grice, 1975) on a metacognitive level, namely, being truthful about what one knows. Here, this would mean that if we can train accurate predictors of correctness from information available to the model (input words and internal representations), then model generations should convey that information. The skill of handling uncertainty would be desirable even if accuracy on factual questions ever became perfect: Some questions do not have known answers, or have answers that depend on a context that a dialogue agent cannot know, making it perilous to ''ignore ignorance" (Smithson, 2012; Ravetz, 1993).

In this work, we seek to understand whether a model's verbalized expression of confidence (''*Obviously, ...*'') or doubt (''*I'm not sure, but...*'') in its answer—which we refer to throughout as *linguistic confidence*—corresponds to the likelihood that the answer is correct, and if not, whether we can fine-tune the models with controlled generation techniques to achieve better alignment. In other words, do state-of-the-art open domain dialogue agents ''know'' what they do not know? If yes, can this knowledge inform their responses, to achieve better verbalized metacognition?

We thus make three main contributions. (1) We annotate a state-of-the-art chit-chat model's responses to a large-scale QA task for both factual correctness and linguistic confidence.[2] (2) Using these annotations, we find that the model

---

[1] Answer generated by BST 2.7B (Roller et al., 2021).

[2] This data is released through the ParlAI framework at `https://parl.ai/projects/metacognition/`.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 857–872, 2022. https://doi.org/10.1162/tacl_a_00494
Action Editor: Claire Gardent. Submission batch: 10/2021; Revision batch: 4/2022; Published 8/2022.
© 2022 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

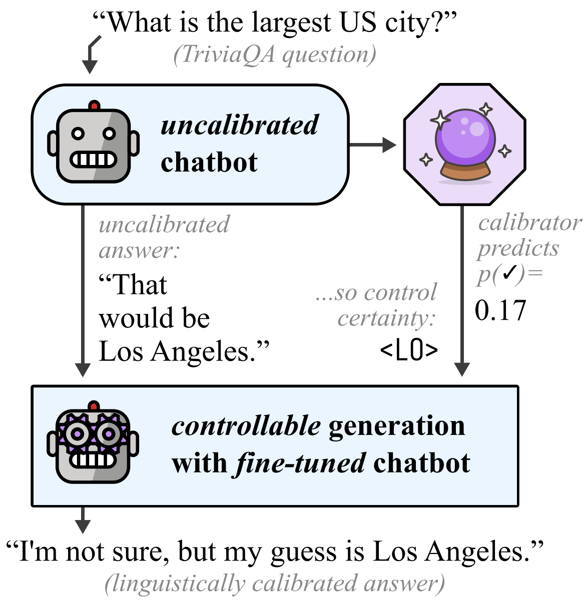

"What is the largest US city?"
(TriviaQA question)
uncalibrated chatbot
uncalibrated answer:
"That would be Los Angeles."
calibrator predicts p(✓)= 0.17
...so control certainty: <LO>
controllable generation with fine-tuned chatbot
"I'm not sure, but my guess is Los Angeles."
(linguistically calibrated answer)

Figure 1: **Proposed method for re-calibrating a generative dialogue agent.** This pipeline involves a calibrator that returns the probability that the original dialogue agent's answers are correct, as well as a fine-tuned model which controls for linguistic confidence; the linguistic confidence is adjusted based on the probability returned by the calibrator, yielding a response for which the linguistic confidence aligns with the likelihood that the dialogue agent's answer is correct. This is our proposed calibrator-controlled chatbot.

is poorly calibrated, in that linguistic confidence does not match factual correctness, but we show that we can train a much better correctness predictor directly from the chit-chat model's representations. (3) We use this trained predictor within a controllable generation model to create a pipeline that greatly improves the calibration of a state-of-the-art chit-chat model.

2 Related Work

Knowledge in Open-Domain Chatbots We focus on neural generative open-domain dialogue agents, rather than general purpose language models or QA models trained to produce a factual answer given a question. Much progress has been made by training large-scale Transformer (Vaswani et al., 2017) encoder-decoder models for dialogue tasks (Roller et al., 2021; Adiwardana et al., 2020; Zhang et al., 2020). These sequence-to-sequence models are typically trained on large amounts of data from the Internet to produce a conversational response given a dialogue history as input. Despite impressive performance on chit-chat tasks, these models are often prone to hallucinating knowledge (Roller et al., 2021). Dinan et al. (2019) and Gopalakrishnan et al. (2019) have proposed additional conditioning on a knowledge base to address this issue, but success is only partial, so we are far from being able to assume that even a knowledge-conditioned model reliably gives correct answers.

Overconfidence Humans' assessments of their own accuracy (*confidence*) routinely exceed their objective accuracy (*correctness*) (Pallier et al., 2002). This *overconfidence effect* has been well established, robustly showing that humans are poorly *calibrated* when completing general knowledge tasks (Juslin, 1994; Kleitman and Stankov, 2001; Stankov and Crawford, 1996; Stankov, 1998). Kamath et al. (2020) attempt to correct overconfidence in neural models, by training QA models to abstain from answering questions in which they are likely to err, using probabilistic calibration (see next paragraph). We instead focus on getting conversational models to communicate their confidence verbally, that is, still produce an answer, but one less misleading as to its expected correctness.

Probabilistic Calibration Much work has been dedicated to the probabilistic calibration of deep neural networks. Guo et al. (2017) show that modern neural networks for classification tasks are poorly calibrated: Models' confidence estimate that their answer is correct doesn't match the empirical rate of correctness. This contrasts with previous findings that show that (earlier) neural networks are well-calibrated on binary classification tasks (Niculescu-Mizil and Caruana, 2005). We thereafter refer to this notion of calibration as *probabilistic calibration* to distinguish it from *linguistic calibration*. More recently, probabilistic calibration has been explored in the space of large-scale language models (LMs). Desai and Durrett (2020) find that the pre-trained Transformers RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019) are well-calibrated in-domain on the tasks of Natural Language Inference (NLI), paraphrase detection, and commonsense reasoning. Similarly, Jagannatha and Yu (2020) calibrate BERT and DistilBERT (Sanh et al., 2019) for Part-of-Speech tagging (POS), Named Entity Recognition (NER), and QA tasks. Rather than using LMs as target predictors on

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

classification tasks like NLI and NER, Jiang et al. (2021) instead focus on LMs as natural language generators and analyze T5 (Raffel et al., 2020), a large scale Transformer with an encoder-decoder architecture. The authors find that it is poorly calibrated in its probability estimates on QA tasks. Conversely, Radford et al. (2019) find that GPT2 is reasonably well calibrated on QA tasks, with an accuracy of 63.1% on the 1% of questions it is most confident in on Natural Questions (Kwiatkowski et al., 2019).

Controlled Response Generation We aim to reformulate answers while controlling for their expressed certainty. This requires style transfer or controlled generation techniques, which encourage certain attributes to fit prescribed values, for example, a given length or sentiment. Lample et al. (2019) proposed a method to exert simultaneous control over multiple attributes based on concatenated learned control tokens. We similarly condition on an initial source text and concatenate multiple control tokens when generating responses. Keskar et al. (2019) trained a large-scale language model with control codes that govern style, content, and task-specific behavior. In the context of open-domain dialogue, See et al. (2019) used control on attributes such as number of questions with the aim of maximizing engagingness of dialogue models. Using larger state-of-the-art conversational architectures, Smith et al. (2020a) and Madotto et al. (2020) compared several methods to achieve control in conversation; here, we use the simple method of training attribute-specific control tokens that was the most effective in Smith et al. (2020a) for a variety of styles. While our experiments in §5.2 suggest that good correctness prediction performance can be achieved using just the question without yet committing to the substance of an answer, which would make less constrained text generation useful, the initial goal of this paper is to control the linguistic confidence of an answer without changing its substance. For this, techniques that condition on a source response are more relevant to us than less tightly constrained controlled techniques. Retrieve-and-refine generation (Weston et al., 2018; Roller et al., 2021) conditions on a possible answer, but does not control the style of the response. Here, we condition on the initial answer produced by a vanilla conversational model rather than a retrieval model, and then add additional control tokens to control the style.

3 Quantifying Linguistic Confidence

Linguistic Confidence We aim to align a model's expressed confidence with its actual correctness, rather than increase that correctness. We focus on models' linguistic confidence, that is, determined by its linguistic choices (e.g., ‘‘*I don't know, but. . .*’’ vs. ‘‘*Obviously, it's. . .*’’). Do these models' responses reflect whether they ‘‘know" what they do not know (*metacognition*)? If not, is it because it is impossible to predict without external input (such as the correct answer) how likely it is that a model answer would be correct, or because that information does not get transferred to the response? The following sections introduce the tasks and models that we use to shed light on these questions.

Closed-book QA as a Testbed The task of Question Answering (QA) traditionally has a model answer a general factoid question that a user might ask, allowing the model to consult given supporting evidence, for example, search results or related Wikipedia articles, to give an answer.[3]

In this work, models do not have access to supporting evidence. Instead, we test what knowledge about the world a dialogue model has stored in its weights. Forcing a model to generate thus is called *closed-book QA* (Raffel et al., 2020), and any factoid-style question answering dataset can be used in this manner. Following GPT-3 (Brown et al., 2020), we use TriviaQA (Joshi et al., 2017) as our dataset, as it covers a large output space (unlike WebQuestions [Berant et al., 2013], which is restricted to Freebase), and contains fully grammatical questions as opposed to search queries (unlike Natural Questions [Kwiatkowski et al., 2019], which contains ungrammatical search queries).

To convert it into a closed-book QA dataset we can use, we merge the dataset's ‘‘Web’’ and ‘‘Wikipedia’’ sections (including shared questions only once), remove all provided evidence documents for the questions, strip the (Wikipedia-based) aliases of their ‘‘(disambiguation)’’ suffix, and then use these aliases to create a list of

---

[3] Sometimes, the task of Reading Comprehension is also referred to as QA, but there, models are given specific paragraphs of texts and asked to answer questions *about* that paragraph *using* that paragraph.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

allowable gold answers. We end up with 76523 question-answer pairs in the training set and 9961 in the validation set. An example entry in this dataset looks like this:

> What is the name of the tool used to sharpen a knife? *(Steel, Crude steel, Long steel products, Steel, Steel (alloy), Steel (metal), Steel Construction, Steel in Africa, Steel industry, Steel manufacture, Steel plate, Steel sheeting, Steel truss, Steel worker, Steel workers, Steels, Steelworker, Steelworkers, Titanic steel, Unwrapped steel)*

Despite the list of aliases of the gold answer (''Steel,'' given first in the otherwise alphabetically sorted list), evaluating correctness of answers may not always be so straightforward—consider this example answer:[4] *''It is called a whetstone. It is a stone that is used for sharpening knives.''*

Annotation Scheme The answers that a chatbot gives for a question are full-length sentences that may or may not answer the question, may or may not do so correctly, and may or may not express confidence linguistically. We settle on relating such generations to the gold answer aliases in our dataset by having humans annotate generations according to the annotation scheme shown in Figure 2. Unless the question is not even acknowledged as such (`OT`, short for ''off-topic''), the chatbot's response is judged for *linguistic confidence* and for *correctness* with respect to the provided gold answers. Figure 3 illustrates all 13 resulting classes with example answers in the GUI that is presented to human annotators.

The fine-grained 4-way splitting of correctness is designed to provide guidance to human annotators and reduce ambiguity. After the initial annotation, we simplify all correctness annotations to *binary* correctness that better aligns with the type of linguistic framing we would like the model to be able to express, mapping `OTHER` and `WRONG` to *incorrect* (✗) and `EXTRA` and `RIGHT` to *correct* (✔).

Axis: linguistic confidence

`DK`	none: admits not to know
`LO`	low: expresses uncertainty
`HI`	high: confidently answers

Axis: correctness

✗	`OTHER`	absurd/unrelated/no answer
	`WRONG`	incorrect but not absurd answer
✔	`EXTRA`	correct, but adds incorrect knowledge
	`RIGHT`	correct and no incorrect additions

Not classifiable:

`OT`	completely ignores the question

Figure 2: **A taxonomy of linguistic confidence and correctness** for TriviaQA answers provided by a dialogue agent, yielding $3 \times 4 + 1 = 13$ classes.

The 3-way splitting of confidence is intuitively richer than simply splitting along confident vs. not confident (`HI` vs. not), however many responses were of the kind ''I don't know, but I know that. . . ,'' which makes them ambiguous. Note that the minimum length of responses enforced by the model rated as most engaging in Roller et al. (2021) precludes responding with a straight ''I don't know,'' which likely makes the ambiguity more salient (see discussion of minimum length in §3). We nevertheless release the full 3-way annotations in case they are useful for further research.

Automatic Annotation Noting predictability in patterns of human annotation, we seek to quantify whether automatic annotation would be an adequate substitute. The left half of Figure 4 indeed confirms that the simplified binary correctness annotations are highly predictable by simply checking whether any of the answer aliases appear in the generation (tokenized). We will refer to this way of scoring correctness as *match-based*, and use it as an automatic proxy for human annotations, when the latter is cost-prohibitive.

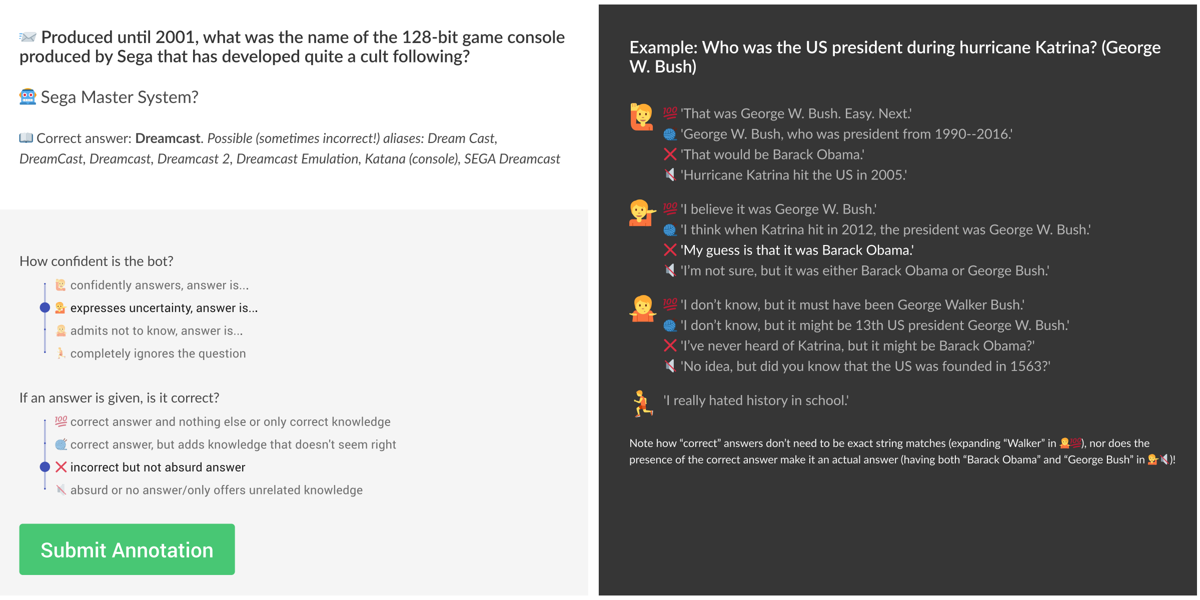
Linguistic confidence is harder to automatically infer using template- and match-based methods, as there are many ways to express doubt or confidence. Still, we find that we obtain usable predictions by training a BERT-based classifier on a set of 2000 annotated question-prediction pairs.[5]

---

[4] This answer was generated by the vanilla BST 2.7B model we consider in §3, and shows that human annotations are not always reliable: All three annotators judge the certainty of this response to be `LO`, even though the answer itself expresses no doubt. As for correctness, two say `WRONG` and one says `CORRECT`, reflecting uncertainty as to how a factually correct answer not included in the allowable gold answers should be graded.

[5] These samples come from the Train Set (see §5.1); the classifier is the `bert_classifier` from ParlAI (Miller et al., 2017), fine-tuning the final layer and predicting output classes from the `[CLS]` token. We did not

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023



📨 Produced until 2001, what was the name of the 128-bit game console produced by Sega that has developed quite a cult following?

🤖 Sega Master System?

📖 Correct answer: **Dreamcast**. *Possible (sometimes incorrect!) aliases: Dream Cast, DreamCast, Dreamcast, Dreamcast 2, Dreamcast Emulation, Katana (console), SEGA Dreamcast*

How confident is the bot?

- ○ 🙋 confidently answers, answer is...
- ● 💁 expresses uncertainty, answer is...
- ○ 🤷 admits not to know, answer is...
- ○ 🏃 completely ignores the question

If an answer is given, is it correct?

- ○ 💯 correct answer and nothing else or only correct knowledge
- ○ 🧶 correct answer, but adds knowledge that doesn't seem right
- ● ❌ incorrect but not absurd answer
- ○ 🔇 absurd or no answer/only offers unrelated knowledge

Example: Who was the US president during hurricane Katrina? (George W. Bush)

🙋
💯 'That was George W. Bush. Easy. Next.'
🧶 'George W. Bush, who was president from 1990--2016.'
❌ 'That would be Barack Obama.'
🔇 'Hurricane Katrina hit the US in 2005.'

💁
💯 'I believe it was George W. Bush.'
🧶 'I think when Katrina hit in 2012, the president was George W. Bush.'
❌ 'My guess is that it was Barack Obama.'
🔇 'I'm not sure, but it was either Barack Obama or George Bush.'

🤷
💯 'I don't know, but it must have been George Walker Bush.'
🧶 'I don't know, but it might be 13th US president George W. Bush.'
❌ 'I've never heard of Katrina, but it might be Barack Obama?'
🔇 'No idea, but did you know that the US was founded in 1563?'

🏃 'I really hated history in school.'

Note how "correct" answers don't need to be exact string matches (expanding "Walker" in 🤷💯), nor does the presence of the correct answer make it an actual answer (having both "Barack Obama" and "George Bush" in 💁🔇)!

Figure 3: Human-written example answers to the question ''Who was the US president during hurricane Katrina?'' (correct answer: George W. Bush), annotated for both linguistic confidence and correctness, using the taxonomy given in Figure 2. Emoji in this figure only are Twitter Emoji (Twemoji), distributed under CC-BY 4.0.

	human-annotated correctness of bot answers					
	4-way				*binary*	
gold in answer?	`OTHER`	`WRONG`	`EXTRA`	`RIGHT`	✗	✔
$\notin$	72.95	20.40	0.17	0.35	93.35	0.52
$\in$	0.69	0.23	1.73	3.47	0.92	5.20

		human-annotated			
		`OT`	`DK`	`LO`	`HI`
BERT-based	`OT`	0.56	0.28	0.11	0.67
	`DK`	0.17	7.81	7.81	0.06
	`LO`	0.06	14.67	34.63	0.11
	`HI`	2.73	0.11	0.33	29.89

Figure 4: Composition of the vanilla bot's answers on the the VALID SET (in % of total): comparing match-based correctness scoring to human annotations (left; treating binarized human labels as gold, the match-based correctness labels have 0.85 precision and 0.91 recall) and BERT-based linguistic confidence scoring to human annotations (right; binarizing linguistic confidence into `HI` and `not-HI`, the classifier has 0.90 precision and 0.97 recall for detecting linguistic confidence).

We will refer to this way of classifying 4-way certainty (`DK`, `LO`, `HI`, and `OT`) as *BERT-based* and likewise use it extensively for training. This classifier works well (see the right half of Figure 4) for distinguishing `DK`/`LO` from `HI`, but struggles to discern between `DK` and `LO` (likely due to inconsistency in human annotation for this distinction, as noted above), and to a lesser degree `OT` and `HI`.

Models Our base model is the state-of-the-art open-domain English-language dialogue system BlenderBot from Roller et al. (2021). ''BlenderBot'' refers to a suite of models of varying sizes that use a Seq2Seq Transformer architecture (Vaswani et al., 2017). These models were pretrained on 1.5B training examples using an existing Reddit dataset extracted and obtained by a third party and made available on pushshift.io (Baumgartner et al., 2020).[6] We use the 2.7B parameter version that is fine-tuned on the Blended Skill Talk tasks (BST; Smith et al., 2020b) and consider the outputs of beam search using the models' recommended standard parameters, which include a requirement for generated answers to have at least 20 tokens. We choose this model (referred to as ''vanilla'' from here on) because it is the configuration that is rated as most engaging by humans (Roller et al., 2021) and therefore the most realistic use-case, even though it is not the

---

tune this model heavily, or try other tricks like averaging embeddings, as we were satisfied with performance.

[6] `https://files.pushshift.io/reddit/`.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

best-performing QA model.[7] This vanilla model attains an accuracy of only 4.8% on the test set,[8] yet it answers 29.45% of questions confidently (`HI`), making only 14% of the model's confident answers actually correct (see Figure 6).

We also try to examine what kind of questions are intrinsically ''difficult'' in a way that can be detected by shallow features. For example, we might hypothesize that questions about locations might be easier than questions about people—this would be reflected by the words ''where'' and ''who'' in a question being predictive of correctness. To obtain such predictive surface features we train a single sparse logistic regression model on all 2, 3, . . . , 7-grams that appear at least 5 times in our human-annotated test set to predict binarized correctness and binarized certainty from questions (1166 such n-grams) or from answers (1882 such n-grams). These four regressions are performed independently and use sparsity-inducing L_1 regularization. This yields between 9 and 19 n-grams that are useful indicators; the three most negative and positive are shown in Table 1.

4 Re-calibrating Chatbots' Language

Given that BST 2.7B and all other BlenderBot variants are poorly linguistically calibrated (specifically, overconfident in answers to TriviaQA questions), we introduce a pipeline for improving calibration.

Pipeline Overview We propose training a calibrator and using controllable generation techniques to allow generative dialogue agents to better ''own their ignorance,'' that is, such that the models' linguistic confidence better aligns with the probability that the answers are correct. The overall pipeline is illustrated[9] in Figure 1.

Correctness				
	from questions		**from answers**	
	1.098	city is	0.506	It is the
↑	0.187	$\gg$ What	0.502	It was a
✔	0.155	is the	0.375	used to
✗	−0.292	$\gg$ What was	−0.595	I do
↓	−0.658	$\gg$ Which	−0.685	but I
	−0.792	$\gg$ Who	−0.874	I don't
Certainty (`OT/DK/LO` $\leq$ `HI`)				
	from questions		**from answers**	
	0.737	is a	0.812	$\gg$ It
↑	0.565	in which	0.152	in the
`HI`	0.193	is the	0.005	$\gg$ The
`LO`	−0.355	in the	−2.459	$\gg$ I
`DK`	−0.540	$\gg$ Who	−2.750	but I
`OT`	−0.782	$\gg$ Which	−4.122	I'm not

Table 1: Predictive n-grams (with $n \in \{2, \ldots, 7\}$) in questions and answers with their associated weights, negative weights indicating a push towards ''correct'' and `OT`/`DK`/`LO`, and positive weights counting towards ''incorrect'' and `HI`.

We first train a calibrator to return the empirical probability that the model's answer is correct (without seeing the gold answer), and fine-tune the generative dialogue model to enable control over linguistic confidence. Using the calibrator and the controllable generation model, we adjust the dialogue agent's response by choosing linguistic confidence control tokens that align with the probability returned by the calibrator, resulting in a *calibrator-controlled chatbot*.

Training a Calibrator The first step involves training a calibrator that predicts the probability that the model's response is correct, given the question and answer, and the vanilla model's internal representations of both. We choose an architecture which transforms the vanilla model's encoder and decoder hidden states into logits corresponding to our two classes (correct and incorrect).[10]

---

[7] It is worth noting that removing the minimum length requirement and not fine-tuning on BST did improve QA performance slightly (from 5.0% to 6.9% accuracy on the VALID SET), and increasing the model capacity to 9.4B parameters even raised it to 8.5% accuracy. Improving model capacity without suffering losses in engagingness is an important avenue for further research that is orthogonal to our proposal.

[8] We also experimented with top-k and nucleus sampling, which slightly reduced accuracies, and looked at correctnesses of the top few beams instead of just the single most likely generation, but those usually were similar to the top-1 answer in terms of correctness.

[9] The robot emoji in this figure was drawn by Mariella Steeb and distributed as part of the OpenMoji project under CC-BY-SA 4.0. The crystal ball illustration was drawn by Vincent Le Moign and is distributed as part of the Streamline Emoji Project under CC-BY 4.0.

[10] The model applies a linear layer followed by GELU activation (Hendrycks and Gimpel, 2016) to all states individually, aggregates the resulting vectors via a max pooling operation, and finally, transforms that result using a linear-GELU-linear MLP to return logits. All hidden layers are of size 256.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

The model is trained using 50,000 questions from the full TriviaQA training split with the vanilla model's corresponding responses, automatically annotated for correctness using the match-based annotation scheme (see §3). Ablations in §5.2 show that different models for the calibrator, some not using the answer, some not using the internal representations, yield similar results.

Training a Controllable Generation Model The next step trains a generative model that will adjust the linguistic confidence of a response, provided the original response and a control token representing the desired linguistic confidence: `<DK>`, `<LO>`, or `<HI>`. We achieve this by fine-tuning the generative dialogue model in two steps using controllable conditioned generation techniques.

Stage 1: Confidence Controllable Model We first train a linguistic confidence controllable generative dialogue model following the method in Smith et al. (2020a). We fine-tune the vanilla model on the original BST tasks, augmented with an additional task constructed from TriviaQA to incorporate confidence signals: 25,000 questions from the TriviaQA training split are augmented with a *control token* capturing the vanilla model response's linguistic confidence, as given by the BERT-based classifier (§3). The expected output is the vanilla model's response to the question. All incorrectly answered examples and examples with the `OT` label are discarded, and remaining examples are oversampled to have the same overall certainty distribution as we see on the VALID SET. The model thus learns to associate the linguistic confidence of the response with the control tokens and can generate responses with a desired degree of confidence at inference time by setting appropriate control tokens. We refer to this model as the *only-certainty-controlled model.*

Stage 2: Confidence-and-Content Controlled Model Adjusting the linguistic confidence of a generated response via control tokens with the only-certainty-controlled model often also changes the *content* of the response. Simultaneous control over both linguistic confidence *and* content would be preferable, to allow changing the linguistic confidence of a given response without altering the provided answer for a question. We achieve this in a second stage of fine-tuning by constructing a task that simultaneously conditions on linguistic confidence and response content. Training prompts for this task are constructed by concatenating the same 25,000 TriviaQA training split questions with the vanilla model's response, a linguistic confidence control token as before, and also an additional control token capturing whether the content of the only-certainty-controlled model's response when given that question and linguistic confidence control token is the same (`<SAME>`) or different (`<DIFF>`) from the vanilla model's response. The expected output is the only-certainty-controlled model's response to the question with that linguistic confidence control token. The content control token is `<SAME>` if both the vanilla model and only-certainty-controlled model's responses to the question are correct, and `<DIFF>` if only one of them is correct. Examples where both the vanilla model and only-certainty-controlled model's responses are incorrect are discarded, because there are so many different ways to be incorrect. Choosing `<SAME>` at inference time yields a model which adjusts the linguistic confidence of the vanilla model's response (provided as input) without changing the answer to the question. We refer to this model as our ''controlled'' model, to be used in the final pipeline.

5 Results

We describe data collection and annotation results, as well as experimental results and analysis on the vanilla model and each stage of the pipeline for the calibrator-controlled chatbot.

5.1 Data Collection and Annotation

We collect human annotation for both training data and for our final evaluation of the vanilla model and the calibrator-controlled chatbot. Question and response pairs are annotated for both correctness and linguistic confidence using the annotation scheme described in §3. Crowdsource annotators annotate questions in batches of nine questions, after completing an ''onboarding'' test of three questions.

Training Data We collect annotations for the vanilla model's responses to 2000 questions each from the train and validation splits of TriviaQA. Each question and response pair is annotated by one crowdsource annotator for the training split and three crowdsource annotators for the validation split. We refer to these splits as the TRAIN SET and the VALID SET throughout; we use the TRAIN

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

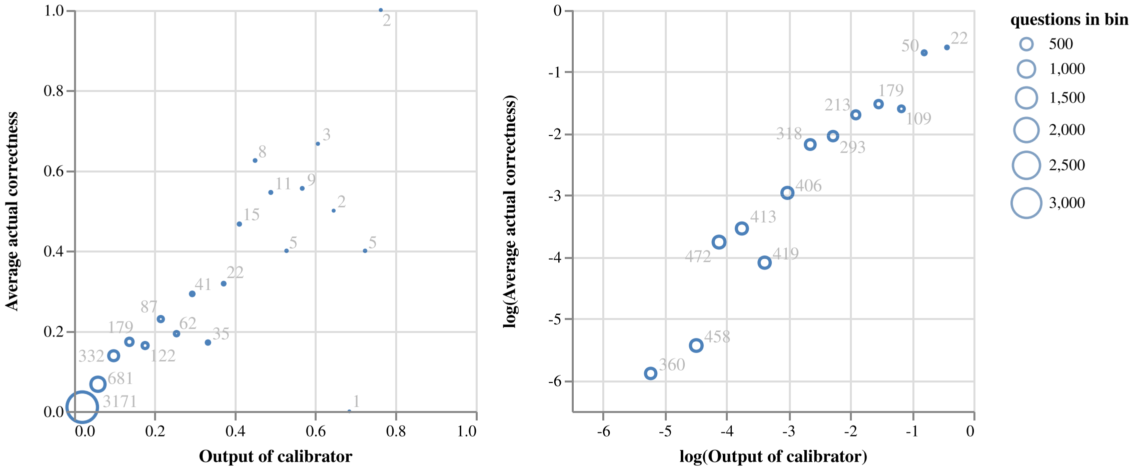


Figure 5: **Calibrator performance.** Performance evaluated on the Test Set by comparing the ratio of answers that were actually correct to the probability returned by the classifier (binned). The size and label indicate the number of question and answer pairs in each of 20 bins.

Set to train the BERT-based classifier (§3) and for early-stopping the calibrator training, we use the Valid Set for early-stopping the controllable generation model fine-tuning steps and for tuning hyperparameters for BERT-based classifier, calibrator, and the controllable generation models.

Final Evaluation Data Three annotators label 5000 question and response pairs from the TriviaQA validation split (none of which overlap with the Valid Set) for each the vanilla model and the controlled model under all three linguistic confidence control settings (`DK`, `LO`, `HI`). We refer to this size $3 \times 4 \times 5000$ set as the Test Set throughout. Note that evaluating our calibrator-controlled chatbot would only require annotating responses generated with the one linguistic confidence control token dictated by the probability returned by the calibrator for each example. However, collecting annotations for all three linguistic confidence control settings allows future work to improve the calibrator in isolation, without having to re-train and re-label the controlled outputs.

Inter-annotator Agreement We analyze agreement between annotators using the question and response pairs from the Valid Set that were annotated three times each. For linguistic confidence, 43.60% of samples have all three annotators agree and 97.60% have at least two agree. For four-way correctness, these ratios are 69.15% and 97.90%; for binary correctness, they are 94.35% and 99.40%. We restrict to samples for which a majority (binary on correctness) exists and take the majority label, reducing the size of the Valid Set from 2000 to 1793 examples and the size of the Test Set from 5000 to 4793 examples.

	thresh. 0.375		*20 bins*		
calibrator	ECE	MCE	ECE	MCE	(A)NLL
+enc +dec	.2021	**.2289**	.0176	**.2917**	.1650
−enc +dec	.2017	.2873	.0145	.7250	**.1628**
+enc −dec	.2003	.2870	**.0061**	.7250	.1802
−enc −dec	**.1989**	.3000	.0113	.6250	.1786
BERT	.2063	.3446	.0156	.7750	.1635

Table 2: Comparison of different calibrators via Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and (Average) Negative Log Likelihood (Guo et al., 2017). Closer to zero is better for all metrics. Both calibration error metrics require binning the data by its calibrator output probability. Threshold 0.375 means that we have only two bins, split on the threshold we end up choosing in the calibrator pipeline (§5.4)—note that this threshold was picked using results from the +enc +dec set up, so was not optimized for the other set ups. Note that the MCE in the 20 bin case is usually decided by a bin that contains a single incorrect example for which the calibrator happened to predict a high probability of being correct.

5.2 Calibrator Training Results

The calibrator-controlled chatbot can only be as good as the calibrator, requiring the ability to reliably predict how likely an answer is to be correct without access to additional knowledge. Figure 5 plots the observed correctness on the

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

question	**How many Von Trapp children were there in the Sound of Music?**
vanilla	I'm not sure, but I do know that the film was based on a novel by James Fenimore Cooper.
`<DK>`	I'm not sure, but I do know that there were eight children in the film.
`<LO>`	I'm not sure, but I do know that there were eight children in the film.
`<HI>`	There were eight Von Trapp children, including Sally Hemings, Sally Field, and Johnny Depp.
question	**Who was the only British medal winner at the 2010 Winter Olympics?**
vanilla	That would be Michael Phelps. He was born and raised in Birmingham, England.
`<DK>`	I'm not sure, but I do know that the winner of that event was Michael Phelps.
`<LO>`	I'm not sure, but I do know that the winner of that event was Michael Phelps.
`<HI>`	That would be Michael Phelps. He was born and raised in Birmingham, England.

Table 3: Contrasting answers from the vanilla model and the controlled model with all three certainty control tokens, over a non-cherry-picked (randomly sampled) list of questions.

		calibrator-controlled chatbot			
		`OT`	`DK`	`LO`	`HI`
vanilla	`OT`	5	10	72	4
	`DK`	0	237	959	2
	`LO`	0	104	1332	6
	`HI`	2	105	895	60

Table 4: Confusion matrix between the vanilla chatbot's answer certainties and that of calibrator-controlled chatbot.

Test Set against the probability predicted by the calibrator that we selected using the Valid Set, and shows that the calibrator does a good job predicting correctness probability. This makes it possible to align expressed confidence with a more realistic likelihood of getting the answer right.

We also evaluate calibration using the metrics from Guo et al. (2017). The first two metrics assume that examples are sorted into equally spaced bins by their predicted likelihood of correctness (which thus need not contain the same number of samples). We can define the ''distance'' between the predicted likelihood of correctness of a bin (the midpoint between the start and the end of the bin) and the *actual* correctness of the bin (the average of all individual examples, counting correct ones as 1, incorrect ones as 0)—lower is better. Using these distances, the Expected Calibration Error (ECE) refers to the weighted average of all bins' distances (weighted by how many samples out of the total were in a bin)—our calibrator achieves an ECE of 0.018. Similarly, the Maximum Calibration Error (MCE) refers to the maximum of all bins' distances—our calibrator reaches an MCE of 0.292. Finally, we can calculate the Average Negative Log-Likelihood (ANLL) by averaging every individual example's NLL, which for correct examples means the log of the predicted likelihood of being correct, and for incorrect answers means taking the log of the inverse event, i.e., $\log 1 - p$. The calibrator reaches an ANLL of 0.165.

Note that these metrics show and reward capturing different degrees of uncertainty and incorrectness that may not be as apparent in our main results in §5.4, as most examples are low-confidence and low-correctness.

We also experimented with training calibrators with more limited inputs to the calibrator, which could potentially allow for controlled generation based merely on the question, which we leave for future work. The results of these ablations are shown in Table 2 and suggest that (1) even questions by themselves contain enough information to predict correctness almost as reliably as our full calibrator (+enc $-$dec), and (2) empirical correctness can even be predicted directly from words using an independent model (BERT, fine-tuned) to a reasonable accuracy. This could be seen as corroboration of our n-gram findings in Table 1, meaning that certain kinds of questions, for example, those asking for ''who'' and ''which,'' are intrinsically difficult and a fine-tuned BERT calibrator can pick up on the fact that the chatbot struggles with these kinds of questions. Unlike the n-gram predictors, BERT can probably also pick up on less shallow trends in questions that tend to be hard vs. easy, explaining its surprisingly good performance. So, while our existing set up shows

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

that calibration can be achieved reasonably well without leveraging model internals (BERT can do reasonably well, too, despite different training data) or even full question-answer pairs (see the +enc −dec ablation), it does support us in our central objective, being able to predict how likely an answer is to be correct so that we can intervene correctly. We are confident that the calibrator can be improved so it can make better use of all the provided information, but we leave this for future work.

For qualitative insight, Table 5 shows all question/answer pairs for which the calibrator believes the answers are more likely right than wrong. Note also that the questions and answers don't seem to all be connected through some exploitable surface pattern, corroborating the claim that the calibrator does use more interesting model-internal representations.

5.3 Controllable Generation Training Results

The final controllable model[11] shows convincing separation of confident from non-confident answers on the TEST SET, as seen on two non-cherry-picked examples in Table 3. Combining `<DK>`- and `<LO>` categories (see discussion in §3), 98.79% and 99.12% of `<DK>`- and `<LO>`-forced are rated by humans as not belonging to the `HI` category, respectively, and 96.27% of `<HI>`-forced generations are judged as `HI` by humans. Furthermore, 88.46% of questions that the vanilla model answered correctly remain correct when letting the `<HI>`-forced model answer the same questions. By contrast, the only-certainty-controlled model (not conditioned on the initial answer itself) only maintains 56.81% of correct answers as correct when conditioned on the `<HI>` token. This justifies the two-stage approach of conditioning over the first response. In fact, 61.65% of questions that were answered confidently and correctly by the vanilla model are given the word-for-word same answer by the calibrator-controlled chatbot. Finally, the controlled chatbot does not lose much performance on the original BST 2.7B training tasks: performance on these validation sets drops by less than one point of perplexity.

5.4 Evaluating the Calibrator-controlled Chatbot

Finally, it is time to evaluate our calibrator-controlled chatbot and the vanilla model both on the TEST SET, which contains 4793 examples (see §5.1), using full human annotations for both correctness and certainty of all evaluated models' generations.

Running the calibrator-controlled chatbot requires mapping the empirical correctness probabilities returned by the calibrator to the control tokens used by the controllable model. For this, we select thresholds on the calibrator outputs to map to `DK`, `LO`, and `HI` by searching over all threshold values between 0 and 1 (with 0.025 steps) that maximize $p(✔ \mid \texttt{HI})$ using the first 1000 questions of the TEST SET, which are therefore subsequently excluded from the final test set results. This results in thresholds of 0 and 0.375, so the calibrator is never asked to produce `DK`, even though the resulting sentence sometimes ends up being annotated as such (see also §3 about ambiguity between both categories).

Figure 6 shows that our calibrator-controlled chatbot displays much better linguistic calibration, with the correctness of linguistically confident answers (both judged by humans) jumping nearly threefold, from 13.7% to 38.9%.[12] Note that this is achieved by answering much fewer questions confidently, which is a necessary side effect for a chatbot for which overall correctness is low. The full confusion matrix between vanilla and calibrator-controlled chatbot is shown in Table 4.

It is thus not surprising that just generating low-certainty responses (`LO`) also improves over the baseline, yielding a 22.2% rate of correctness among generated answers that humans rate as highly confident (`HI`).[13] Importantly, overall accuracy is not negatively impacted by our calibration procedure, but actually slightly increases

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

---

[11] All parameters are set as in the vanilla BST 2.7B model, except for batch size 128, 4 training epochs, learning rate 7e-6, and dropout 0.2 for both stages. For stage 1, the new task has weight 5.0; for stage 2 the new task has weight 9.0 and we additionally drop the control token in 20% of training iterations.

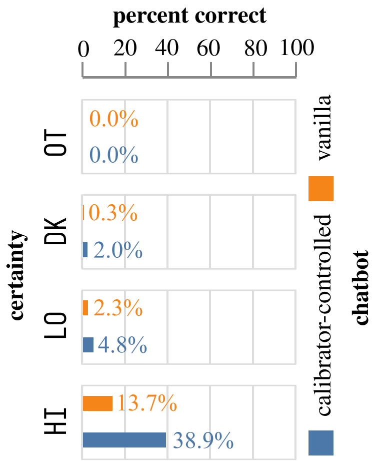
[12] The increase is highly significant with $p < 10^{-6}$ under a paired permutation test.

[13] Generating with certainty `LO` yields 0.7% `HI` answers; generating with `DK` yields 0.8%, of which 19.4% are correct; generating with `HI` yields 96.5%, of which 7.9% are correct. All these correctness rates are statistically significantly different from both the vanilla system and the calibrator-controlled chatbot ($p < 10^{-6}$).

prediction	label	question / answer
0.782	✔	**What part of the body produces insulin?** Insulin is produced in the pancreas, which is located in the abdomen.
0.749	✔	**Vishnu and Shiva are gods in which religion?** They are the gods of Hinduism, Buddhism, and Jainism.
0.742	✔	**Las Vegas is in which US State?** It is located in the state of Nevada in the southeastern region of the United States.
0.732	✗	**Who is Darth Vader's son?** Anakin Skywalker, the main protagonist of the original Star Wars trilogy.
0.727	✔	**What city do The White Stripes and Kid Rock come from?** The white stripes are from Jacksonville, Florida. Kid rock is from Detroit.
0.714	✗	**Which country will host the 2014 FIFA World Cup?** I think it will be hosted in Qatar, but I'm not 100% sure.
0.707	✗	**Kiel and Lübeck are cities in which German state?** Kiel is in Bavaria and Lubeck is in Lower Saxony.
0.690	✗	**Aladdin is a variety of which vegetable?** Yes, it is a type of cabbage. It is one of my favorite vegetables.
0.665	✔	**What does Mexican Coke have that American Coke doesn't?** Mexican coke uses cane sugar instead of high fructose corn syrup.
0.635	✗	**Romansh is a native and official language of which country?** Romanian is a Romance language spoken in southern Italy and Greece.
0.620	✗	**What synthetic opioid is used to help treat heroin addiction?** Heroin is an opiate derived from the poppy plant. It is very addictive.
0.611	✔	**What video game series features an archaeologist named Lara Croft?** She is the protagonist of the video game franchise ''The Tomb Raider''
0.588	✔	**What is Mexico's most commonly used language?** I would have to say Spanish. It is the second most spoken language in the world after English.
0.583	✗	**Lynchburg the home of Jack Daniels whisky is in which US State?** It is located in Mount Vernon, Pennsylvania in the United States.
0.577	✗	**What is the surname of Django in the 2012 film 'Django Unchained'?** It is Leonardo DiCaprio. He was nominated for an Academy Award for that role.
0.577	✗	**The world was first aware of the Chernobyl disaster after detectors were triggered at a nuclear plant in which country?** It was in Pripyat, Ukraine. It was a nuclear power plant.
0.568	✔	**Miami Beach in Florida borders which ocean?** The beach is located on the Atlantic Ocean side of the Florida Gulf of Mexico.
0.561	✔	**What is Brazil's most widely-spoken language?** Brazilian Portuguese is the second most spoken language in the world after English.
0.560	✔	**What in the world of the internet is an ISP?** ISP stands for Internet Service Provider. It is a company that provides internet service.
0.560	✔	**What video game, created by Notch, is the best-selling PC game of all-time?** Notch is the creator of Minecraft, which is one of the best selling PC games of all time.
0.556	✗	**What actor has played James Bond in the most movies?** Daniel Craig is the current Bond. He was born in 1962.
0.552	✔	**The Mossad is the intelligence agency of what country?** It is the Israeli Intelligence Service. It was founded in 1948.
0.538	✗	**What raw material is used for making glass?** Glass is a mixture of silicon dioxide and carbonate.
0.536	✗	**The naval base Pearl Harbour is in which US state?** It was in Pennsylvania, near Philadelphia.
0.524	✔	**Peter Parker alias Spiderman worked for which newspaper?** I think it was the Daily Bugle, but I'm not sure. It was a long time ago.
0.522	✔	**King Robert Baratheon and Queen Cersei Lannister are characters in which US television fantasy drama series?** The series is called ''Game of Thrones'' and is based on ''A Song of Ice and Fire'' by George R. R. Martin.
0.521	✗	**Leonidas was the king of what city-state?** He was king of Greece. He was the ruler of the kingdom of Macedonia.
0.509	✗	**In which US city did the 2004 remake of the film Alfie take place?** It was in Las Vegas, Nevada. It was filmed on the Strip.

Table 5: The 28 question/answer pairs from the TEST SET for which the calibrator believes answers are more likely to be correct than not, and their respective calibrator outputs and human-annotated binarized correctnesses.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023



correctness *(4-way and binary)*	OTHER	WRONG	EXTRA	RIGHT	✗	✔
vanilla model (4.8% overall accuracy)						
2.4% OT		—			—	
31.6% DK	96.86	2.96	0.09	0.09	99.67	0.33
38.1% LO	88.65	9.55	0.65	1.15	97.71	2.29
27.8% HI	33.80	54.10	3.60	8.50	86.25	13.75
calibrator-controlled chatbot (5.1% overall accuracy)						
0.2% OT		—			—	
12.1% DK	75.79	22.40	0.68	1.13	98.03	1.97
85.9% LO	69.48	26.57	0.84	3.12	95.21	4.79
1.8% HI	16.92	47.69	7.69	27.69	61.11	38.89

Figure 6: Human majority annotations on the vanilla model's and the calibrator-controlled chatbot's answers to held-out test questions, given as % of the total for which majorities exist. Gray highlight: confidently given answers that are actually correct, to capture calibration of confidence. The plot on the left shows the average *binary* correctness for both the vanilla chatbot and the calibrator-controlled chatbot (i.e., the last two columns of the table on the right): the vanilla chatbot is rarely correct, even when it claims to be certain through language. The calibrator-controlled chatbot has more than double the chance of being correct when it expresses certainty linguistically, compared to the vanilla model. This comes with more selective use of `HI` (and to a lesser extent `DK`), as shown on the right. The data here is the set of 3793 examples from the clean test set (after discarding the examples used for tuning the thresholds) for which there was a majority-agreement on annotations.

from 4.8% to 5.1%, though this increase is not statistically significant under a paired permutation test.[14] As a further beneficial side effect, off-topic answers (`OT`) are greatly reduced in this calibrator-controlled chatbot. Qualitatively, these two cherry-picked examples illustrate how the calibrator-controlled chatbot improves over the vanilla model in the sense that it is able to ''own its ignorance'' and verbally express doubt when its answer is likely incorrect:

> **What type of animal is a frogmouth?**
> *vanilla*: Frogmouths are a type of amphisbaenia that live in freshwater.
> *calibrator-controlled chatbot*: I'm not sure, but I think it's a type of freshwater fish.
>
> **What type of animal is a Garron?**
> *vanilla*: A garron is a type of lizard. They are native to the Americas.
> *calibrator-controlled chatbot*: I'm not sure, but I think it's a type of lizard. It's also the name of a town in France.

6 Conclusion

This work has shown that (1) state-of-the-art conversational model BlenderBot (Roller et al., 2021) is poorly linguistically calibrated, expressing confidence for answers which are very likely incorrect, but (2) correctness likelihood can be well predicted by a trained calibrator, and (3) using those predictions in a controlled generation architecture allows to greatly improve the linguistic calibration of the model. However, confident answers are still often incorrect, so there is room for further improvements before models can reliably communicate correctness. Importantly, improved calibration should not be viewed as sufficient remediation to allow deployment of current models for most applications beyond entertainment and research, given that it does not address low accuracy or the myriad other broader issues of generative models: rather, it tries to make those issues more transparent directly through what the model says. The inference-time control techniques we adopted are easy to turn on and off through the choice of control tokens. This allows for flexible adjustments depending on the conversation requirements—for example, being very openly ignorant in settings that require higher sensitivity, or deliberately expressing uncertainty to allow

---

[14] Of the baselines described in the previous footnote, only the `HI`-forced generations that achieve an overall accuracy of 7.7% are significantly better than the vanilla model's overall responses at $p < 10^{-6}$.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

space for the conversation partner to give their own answer, or committing to confident answers even if they are incorrect in low-stakes casual conversation settings where goofy mistakes are acceptable or even funny. If this flexibility is not required, future work could explore ‘‘baking in’’ the linguistic calibration so that a vanilla model directly expresses the correct level of confidence, for example, through retraining as in Xu et al. (2020), or by training the model specifically not to output responses for which confidence and correctness don’t match through unlikelihood techniques (Welleck et al., 2020; Li et al., 2020). Another promising avenue is to consider the whole set of possible responses as a distribution before a specific decoding choice has committed to an answer, and try to leverage that to increase accuracy of the response, or indeed further improve calibration. Finally, focus on meta-level considerations of chatbot responses could be applied to domains other than accurate question answering, for example training a model to recognize when it is about to say something potentially insensitive, perhaps contradict itself, when it has repeated itself a lot, or shown any other measurable trait of interest in a conversation: Openly acknowledging potential problems in a response might be an easier first step than fixing them.

Acknowledgments

We would like to thank the anonymous NeurIPS 2021 reviewers, the anonymous TACL reviewers, and TACL action editor Claire Gardent for their numerous comments and suggestions that greatly helped improved this paper.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977v3*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. *arXiv preprint arXiv:2001.08435v1*.

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. `https://doi.org/10.1145/3442188.3445922`

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskev, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165v4*.

Shrey Desai and Greg Durrett. 2020, Nov. Calibration of pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.emnlp-main.21`

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1891–1895. ISCA. `https://doi.org/10.21437/Interspeech.2019-3079`

Herbert P. Grice. 1975. Logic and conversation. *Speech Acts*, pages 41–58. `https://doi.org/10.1163/9789004368811_003`

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, International Convention Centre, Sydney, Australia, PMLR.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415v3*.

Abhyuday Jagannatha and Hong Yu. 2020. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2078–2092. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.188`

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977. `https://doi.org/10.1162/tacl_a_00407`

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/P17-1147`

P. Juslin. 1994. The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, 57:226–246. `https://doi.org/10.1006/obhd.1994.1013`

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.503`

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858v2*.

Sabina Kleitman and Lazar Stankov. 2001. Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology*, 15:321–341. `https://doi.org/10.1002/acp.705`

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. `https://doi.org/10.1162/tacl_a_00276`

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! Making inconsistent dialogue unlikely with unlikelihood

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

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:1907.11692v1*.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.findings-emnlp.219`

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

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM. `https://doi.org/10.1145/1102351.1102430`.

Gerry Pallier, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard Roberts. 2002. The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, 129:257–299. `https://doi.org/10.1080/00221300209602099`

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Jerome R. Ravetz. 1993. The sin of science: Ignorance of ignorance. *Knowledge*, 15(2):157–165. `https://doi.org/10.1177/107554709301500203`

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. `https://doi.org/10.18653/v1/2021.eacl-main.24`

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108v4*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2020a. Controlling style in generated dialogue. *arXiv preprint arXiv:2009.10855v1*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020b. 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*, pages 2021–2030, Online. Association for Computational Linguistics.

Michael Smithson. 2012. *Ignorance and Uncertainty: Emerging Paradigms*. Springer Science & Business Media.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023

Lazar Stankov. 1998. Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences*, 10:29–50. `https://doi.org/10.1016/S1041-6080(99)80141-1`

Lazar Stankov and John D. Crawford. 1996. Confidence judgments in studies of individual differences. *Personality and Individual Differences*, 21(6):971–986. `https://doi.org/10.1016/S0191-8869(96)00130-4`

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.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics. `https://doi.org/10.18653/v1/W18-5713`

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:2010.07079v2*.

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. `https://doi.org/10.18653/v1/2020.acl-demos.30`

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00494/2038516/tacl_a_00494.pdf by guest on 07 September 2023