Probabilities of Chat LLMs Are Miscalibrated but Still Predict Correctness on Multiple-Choice Q&A

Anonymous authors
Paper under double-blind review

Abstract

We study 15 large language models (LLMs) fine-tuned for chat and find that their maximum softmax probabilities (MSPs) are consistently miscalibrated on multiple-choice Q&A. However, those MSPs might still encode useful uncertainty information. Specifically, we hypothesized that wrong answers would be associated with smaller MSPs compared to correct answers. Via rigorous statistical testing, we show that this hypothesis holds for models which perform well on the underlying Q&A task. We also find a strong direct correlation between Q&A accuracy and MSP correctness prediction, while finding no correlation between Q&A accuracy and calibration error. This suggests that within the current fine-tuning paradigm, we can expect correctness prediction but not calibration to improve as LLM capabilities progress. To demonstrate the utility of correctness prediction, we show that when models have the option to abstain, performance can be improved by selectively abstaining based on the MSP of the initial model response, using only a small amount of labeled data to choose the MSP threshold.

1 Introduction

Large language models (LLMs) have demonstrated profound capabilities in many domains, but continue to sometimes generate plausible-sounding false responses (Huang et al., 2023). In one high-profile case, an LLM-based system invented a litany of nonexistent court cases, leading to formal sanctions for two lawyers (Mangan, 2023). Although ongoing work has reduced the rate of these mistakes,[1] LLMs will inevitably face situations that surpass the boundaries of their existing knowledge. In those situations, it is unrealistic to expect these models (or any intelligent agents, including humans) to always make perfect decisions. Rather than confidently misleading users, LLMs should be able to detect unfamiliar situations and act cautiously (e.g., decline to answer).

In this paper, we study whether LLMs can determine the correctness of their own answers to multiple-choice questions. If so, this would directly enable LLMs to decline to answer when they are likely to be incorrect. Prior work has fine-tuned LLMs with labeled data to recognize questions beyond their knowledge (Kadavath et al., 2022; Zhang et al., 2023), but this approach is costly and could interfere with other capabilities of the model. More broadly, such approaches do not answer the fundamental scientific question of whether LLMs already possess the necessary information to decide when to abstain. That is, can this ability be *invoked* rather than *learned*?

We investigate the maximum softmax probability (MSP) as a potential source of innate uncertainty information. Specifically, our goal is to understand the relationship between the MSP of an LLM response and the correctness of that response.

We evaluate 15 LLMs fine-tuned for chat (henceforth "chat LLMs" for brevity) on five different Q&A datasets. Figure 1 shows a sample question prompt. The selected LLMs cover a range of sizes, capabilities and architectures, and include both open-source and proprietary models. To our knowledge, our work is the most comprehensive study on LLM correctness-awareness. This comprehensiveness bolsters the robustness of our claims but also enables *cross-model* comparisons which reveal novel insights.

[1] See, for example, https://huggingface.co/spaces/hallucinations-leaderboard/leaderboard.

```
Below is a multiple-choice question.  Choose the letter which best answers
the question.  Keep your response as brief as possible; just state the letter
corresponding to your answer, followed by a period, with no explanation.

Question:
In the nitrogen cycle, nitrogen can return to the lithosphere directly from the
atmosphere by
A. lightning.
B. cellular respiration.
C. air pollution.
D. condensation.

Response:
```

Figure 1: A sample question prompt.

Calibration. We first ask whether the MSP is *calibrated* (DeGroot & Fienberg, 1983; Nguyen & O'Connor, 2015), meaning that among responses with an MSP of $p\%$, $p\%$ are correct. Calibrated MSPs enable fully unsupervised abstention policies with theoretical guarantees: a calibrated model that answers only when the MSP is higher than $1 - \varepsilon$ guarantees that the chance of an incorrect answer is at most ε. However, we show that this approach is not generally viable for chat LLMs. Figure 2 (left) shows that most LLMs in our study have poorly calibrated MSPs. In particular, the MSPs are consistently overconfident. Furthermore, improved performance on the underlying Q&A task does not necessarily translate to better calibration: Figure 2 (right) shows no correlation between overall Q&A accuracy (percentage of questions answered correctly) and calibration error ($p = 0.32$). This is one of the cross-model comparisons mentioned above.

Correctness prediction without calibration. Even if the MSP cannot be directly interpreted as the probability of correctness, it might still be predictive of correctness. As a simplified example, consider a model whose MSP is consistently 0.9 for correct responses and 0.8 for incorrect responses. This model is clearly miscalibrated, but its MSP perfectly predicts correctness.

Through rigorous statistical testing, we demonstrate that the MSPs of chat LLMs can indeed predict correctness.[2] Moreover, the predictiveness is stronger for models which perform better on the underlying Q&A task ($p < 10^{-3}$). This second cross-model comparison suggests that the ability to predict correctness will strengthen as the general capabilities of LLMs improve (e.g., by scaling up data and model sizes). The same is not true of calibration, as discussed above. These contrasting results reveal a novel dichotomy between two approaches to uncertainty quantification.

Q&A with abstention. In addition to demonstrating the predictive power of the MSP and maximum logit, we provide a proof-of-concept for how this information can be leveraged to reduce LLM harm in practice. We analyzed a variant of the original Q&A task where models can also abstain and receive 1 point per correct answer, 0 points per abstention, and $-c$ points per wrong answer. We found that for both $c = 1$ and $c = 2$, selectively abstaining based on the MSP and/or maximum logit led to substantial improvements over the base models. We used a mere 20 data points (i.e., 20 randomly selected questions and their answers) to select the abstention threshold.

In summary, we show the following:

1. The MSPs of chat LLMs are miscalibrated, and this does not improve as model capabilities improve.
2. The MSPs of chat LLMs still predict correctness, and this *does* improve as model capabilities improve.
3. A small amount of labeled data is enough to translate correctness prediction into an effective abstention method.

[2]Correctness prediction is measured by the Area Under the Receiver Operating Characteristic curve (AUROC). See Section 5 for details.

The paper proceeds as follows. Section 2 discusses related work and Section 3 covers our general experimental setup. Section 4, 5, and 6 present our calibration results, AUROC results, and Q&A-with-abstention results, respectively.

2 Related work

Key comparison: Kadavath et al. (2022). The most relevant prior paper is Kadavath et al. (2022), who studied LLM correctness-awareness in a variety of domains. There are three key differences between their work and ours.

The first is that *they primarily studied raw pretrained (i.e., not fine-tuned) LLMs.* In particular, their well-known finding that LLMs are well-calibrated applies only to raw pretrained models. In this way, our work complements theirs by showing that their finding of good calibration fails to generalize to LLMs fine-tuned for chat. (They actually briefly analyzed one fine-tuned model and found that it is quite miscalibrated.) Correctness-awareness may also be more important for fine-tuned models, since the casual user is less likely to interact with raw pretrained LLMs.

The second is that *they only studied MSP calibration and not MSP correctness prediction.*[3] This may be because good calibration directly yields theoretically grounded correctness prediction, so for raw pretrained models, good calibration may suffice. However, our finding that chat LLMs have miscalibrated MSPs motivates the separate question of whether MSPs can predict correctness.

The third is *comprehensiveness.* They only tested a single series of models, while we test 6 series of models (or 8, depending on how one counts) and 15 models total. Our comprehensiveness crucially enables cross-model comparisons, as discussed in Section 1. In particular, we have statistical evidence that the correlation between correctness prediction and Q&A accuracy (and the lack of a correlation between calibration and Q&A accuracy) may extend to models that do not even exist yet. In contrast, it is harder to claim that the findings of Kadavath et al. (2022) generalize to other models, since they essentially have a sample size of one.

Overall, our work complements theirs. Viewing our work and theirs side-by-side suggests that fine-tuning degrades calibration of LLMs and this effect is not mitigated as models become more capable. However, this procedure only *distorts* rather than erases uncertainty information in LLMs, and that uncertainty information *does* become more useful as models become more capable.

LLM calibration. Uncertainty quantification in LLMs is a very active area and a full survey is beyond the scope of this paper; we direct the interested reader to Geng et al. (2024). To our knowledge, the most relevant prior papers on MSP calibration are OpenAI (2023) and Zhu et al. (2023). OpenAI (2023) provides a thorough calibration study of GPT-4 and, similar to our work, finds that the MSPs are consistently overconfident. Zhu et al. (2023) studies calibration for both fine-tuned and non-fine-tuned models, but their study of fine-tuned models is limited to Llama 7B. We argue that robust conclusions on MSP calibration cannot be drawn from two separate studies of a single model each. In contrast, we present a comprehensive and unified evaluation of the MSP calibration of 15 LLMs fine-tuned for chat.

Training LLMs to abstain. Another line of work has fine-tuned LLMs to predict the correctness of their answers (Kadavath et al., 2022; Yin et al., 2023; Zhang et al., 2023). This approach has a different focus than our work: our goal is to understand the fundamental relationship between MSPs and correctness, not to design a state-of-the-art abstention method. Our Q&A-with-abstention experiments are intended as a simple proof of concept of our correctness prediction findings. In other words, our primary contribution is scientific, not methodological.

Abstaining based on the MSP. The idea of abstaining based on the MSP was originally introduced by Chow (1970) in the context of pattern recognition. This technique was recently explored for LLMs by Gupta et al. (2024), although their setting is different. Also, their experiments only use FLAN-T5 models. In contrast, we test 15 different LLMs, which enables the cross-model comparisons previously discussed.

[3]They did study correctness prediction in the different context of training LLMs to abstain. We discuss those results separately below.

Beyond LLMs. The MSP has been used for anomaly/out-of-distribution detection in a variety of other contexts, including coreference resolution tasks (Nguyen & O'Connor, 2015) pre-trained BERT models (Hendrycks et al., 2020), and image classification (Hendrycks et al., 2022; Hendrycks & Gimpel, 2017).

3 Experimental setup

This section presents our general experimental setup. Elements of the setup which are specific to calibration, correctness prediction, or abstention are discussed in Sections 4, 5 and 6, respectively. We have attached the code for all of our experiments and analysis as supplementary material.

Multiple-choice Q&A. We chose to study multiple-choice Q&A because there is exactly one correct answer: this allows us to study the core hypothesis of whether MSPs are predictive of correctness without having to deal with more complex issues such as degrees of correctness or multiple valid phrasings of the same correct answer.

Datasets. We based our experimental framework on the original Hugging Face LLM leaderboard (Beeching et al., 2023). We used all five multiple-choice Q&A datasets[4] from that leaderboard: ARC-Challenge (Clark et al.), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2021). We randomly sampled 6,000 questions from each dataset, except for those with fewer than 6,000, in which case we used all of the questions (ARC-Challenge and TruthfulQA have 2,590 and 817 questions, respectively). The 6,000 number was chosen to make the experiment duration manageable.

Prompting style. To test our hypothesis in the simplest possible setting, we used a plain zero-shot prompting style. For comparison, we also ran the experiments with 5-shot prompting and obtained similar results (Appendix B). We also used two different phrasings to ensure that our results were not an artifact of the specific phrasing. One phrasing is shown in Figure 1 and the other appears in Appendix A.

Models. We tested 15 LLMs: 13 open-source and two proprietary. The open-source LLMs were chosen based on a combination of performance on the aforementioned leaderboard and the number of downloads on Hugging Face. The open-source models we selected are Falcon (7B and 40B) (Almazrouei et al., 2023), Llama 2 (7B, 70B) (Touvron et al., 2023), Llama 3 (8B, 70B) and Llama 3.1 (8B, 70B) (AI@Meta, 2024), Mistral 7B v0.2 (Jiang et al., 2023), Mixtral 8x7B (Jiang et al., 2024), SOLAR 10.7B (Kim et al., 2023), and Yi (6B and 34B) (01-ai, 2023). All of the open-source LLMs were accessed through the Hugging Face interface and were run with dynamic 4-bit quantization, which has been shown to preserve performance while reducing computational requirements (Dettmers et al., 2023). The experiments on open-source LLMs took about 1000 GPU-hours using NVIDIA RTX A6000 GPUs.

We used the fine-tuned "chat" or "instruct" versions of all models. Our focus on fine-tuned models is due to the unreliability of the non-fine-tuned base models. We were unable to get many of the base models to consistently respond in the correct format, despite significant prompting effort. As discussed in Section 2, Kadavath et al. (2022) do study the calibration of base models, but they only study a single family of models. Our guess is that that particular family of base models is more reliable in terms of response format.

We also tested two proprietary LLMs for which we could obtain softmax probabilities through an API: OpenAI's GPT-3.5 Turbo (Brown et al., 2020; Ouyang et al., 2022) and GPT-4o (OpenAI; 2023). The OpenAI API does not provide pre-softmax logits, so we could not compute Max Logit AUROCs for those two models.[5] These experiments took about a day and cost about $120.

Aggregating results across datasets. Grouping the questions from all datasets together to compute a single AUROC per model would undervalue datasets with fewer questions. Instead, we computed a separate AUROC for each available combination of model, dataset, prompt phrasing, and classifier (MSP vs Max Logit).[6] All in all, we recorded 280 AUROC data points (15 models $\times$ 5 datasets $\times$ 2 phrasings $\times$ 2

[4]The leaderboard also includes the GSM8k dataset, which we excluded since it is not multiple-choice.

[5]We also could not test "reasoning" models such as o1 and o3-mini, since the API provides neither logits nor probabilities for those models.

[6]AUROC was computed by the Python module sklearn.

classifiers, excluding Max Logit for OpenAI models) and 150 Q&A accuracy data points (15 models $\times$ 5 datasets $\times$ 2 phrasings) over a total of 642,210 prompts (15 models $\times$ 21,407 questions across datasets $\times$ 2 phrasings). We then calculated per-model unweighted averages to get the results in Table 1.

3.1 Computing MSP and Max Logit

Let V be a set of tokens (a vocabulary) and let $\boldsymbol{x}$ be a sequence of tokens from V. For each token $y \in V$ and prefix $\boldsymbol{x}$, an LLM computes a logit $L(y; \boldsymbol{x})$. A softmax function is applied to the logits to derive the probability of y being the next token:

$$P(y \mid \boldsymbol{x}) = \frac{\exp(L(y; \boldsymbol{x}))}{\sum_{z \in V} \exp(L(z; \boldsymbol{x}))}$$

The Max Logit of token y is $\max_{y \in V} L(y; \boldsymbol{x})$ and the MSP is $\max_{y \in V} P(y \mid \boldsymbol{x})$. Note that the Max Logit and MSP both correspond to the same token: $\arg\max_{y \in V} L(y; \boldsymbol{x}) = \arg\max_{y \in V} P(y \mid \boldsymbol{x})$. In our experiments, we formulated each question as a prompt (Figure 1), then used greedy decoding to generate a response (i.e., we always picked the token with the maximum logit).

For typical supervised learning classification tasks such as those studied in Hendrycks & Gimpel (2017) and Hendrycks et al. (2022), there is a single MSP and maximum logit per response. However, an LLM response can consist of multiple tokens and the model produces an MSP and Max Logit for *each token*.

Ideally, we would like to extract from the response a single token that indicates the LLM's answer to the question. For this reason, our prompts state that the models should reply in the format "A."/"B."/etc. Fortunately, nearly all responses were in the correct format: 87% of responses started with "A."/"B."/etc, and 92% contained "A."/"B."/etc somewhere in the response. Those numbers increase to 93% and 97% respectively when the weaker models are excluded (Falcon, Llama 2, Mistral, Yi 6B).

As such, we simply searched the output string for the first occurrence of "A."/"B."/etc and then recorded the MSP and Max Logit corresponding to that capital letter token, since that capital letter token determines the LLM's answer. If there was no occurrence of "A."/"B."/etc , we searched for just "A"/"B"/etc, although as mentioned, this case was rare. The same search process was used for computing MSP/Max Logit and evaluating correctness. If the search failed, we recorded the MSP and Max Logit as zero, and labeled the LLM's response as wrong.[7] To verify that our results are not sensitive to the specifics of this search process, we performed a secondary analysis where we replaced the MSP of the answer token with the product of MSPs across all tokens. Those results were similar and can be found in Appendix C.

4 Calibration

We first asked whether the MSPs of chat LLMs are calibrated. For each possible combination of LLM, dataset, and prompt phrasing, we performed two analyses. First, we computed each model's calibration curve as follows. For each model we divided the range of possible MSPs into 10 quantile bins, i.e., each containing the same number of data points. Then for each bin, we computed the average MSP and the fraction of correct responses in that bin. Figure 2 (left) displays each model's calibration curve. Most models exhibit clear overconfidence: the MSP is consistently much larger than the fraction of correct responses. For example, most models produce MSPs above 0.95 even when they are correct only 60% of the time.

This shows that most models are miscalibrated, but how does the level of miscalibration vary between models? To answer this question, we computed each model's total calibration error, equal to the mean over bins of the absolute difference between the average MSP and the fraction of correct answers. To avoid downweighting smaller datasets, we first computed the calibration error for each model-dataset pair and then computed per-model unweighted averages across the five datasets (and two prompt phrasings). We then plot each model's calibration error vs its overall Q&A accuracy. This allows us to see whether calibration error goes down (or up) as models become more powerful.

[7] We considered excluding these responses, but we reasoned that practically speaking, they are incorrect. This was also not an impactful decision, as only 0.4% of responses were unparseable. This number drops to 0.1% when the Falcon, Llama 2, Mistral, and Yi 6B models are excluded.

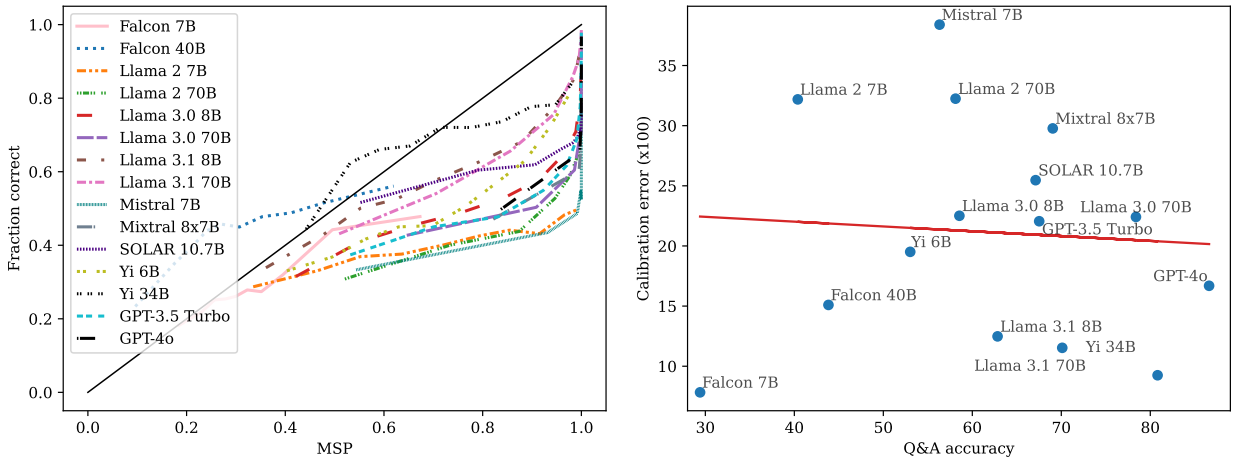

Figure 2: **Left:** The calibration curve for each model. Most models exhibit clear overconfidence: the MSP is larger than the true fraction of correct responses. **Right:** The data from Figure 2 is aggregated to a single data point per model: the model's total calibration error (scaled by 100 to improve readability). There is no statistical evidence for a correlation between calibration error and Q&A accuracy ($r = -0.07, p = 0.81$). The precise calibration error values can be found in the appendix (Table 4).

Figure 2 (right) shows that the answer is no. Although a slight downward trend exists, it is mostly due to three outlier models: Yi 34B and Llama 3.1 8B and 70B.[8] More formally, the Pearson correlation coefficient (PCC) was $r = -0.07$ with $p = 0.81$, indicating no correlation.[9] As such, we should not expect chat LLMs to become more calibrated as their capabilities grow. Despite this, we will see in Section 5 that we *can* expect MSPs to become increasingly effective at predicting correctness, even as their calibration does not improve.

5 Predicting answer correctness with MSP and Max Logit

Before presenting said results, we must define correctness prediction. Correctness prediction is a binary classification task: given a multiple-choice question and the LLM's response, predict whether the response is correct. We study two classifiers for this task: the MSP classifier (predict correctness iff the MSP exceeds some threshold) and the Max Logit[10] classifier (predict correctness iff the maximum pre-softmax logit exceeds some threshold). We hypothesized that the MSP and Max Logit classifiers would (statistically) outperform random chance on this classification task.

Note that we are not training a new binary classifier for this task: we study the performance of the MSP and Max Logit "out of the box", since our goal is to understand the innate properties of LLMs.

5.1 Methodology

AUROC. Performance on a binary classification task is often measured by the Area Under the Receiver Operating Characteristic curve (AUROC) (Bradley, 1997). The AUROC of a binary classifier ranges from 0% to 100%, where 0% corresponds to getting every prediction wrong, 50% is random chance, and 100% is perfect classification. AUROC is also equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. Conveniently, AUROC is threshold-independent in that it captures the model's performance across the entire range of possible thresholds.[11]

[8] Interestingly, the calibration error of the Llama 3.1 models is much lower than that of their 3.0 predecessors, despite the Q&A accuracy being very similar in the two families.

[9] PCCs and p values for cross-model correlations (e.g, Figures 2 (right) and 3) were computed via standard linear regression.

[10] Max Logit has no notion of calibration, since it is not a probability.

[11] Note that calibration error is also threshold independent.

We computed the AUROC for each available combination of LLM, dataset, prompt phrasing, and classifier (MSP or Max Logit). We could not compute Max Logit AUROCs for the OpenAI models (GPT-3.5 Turbo and GPT-4o) because the OpenAI API only provides softmax probabilities and not pre-softmax logits. This resulted in 280 AUROC data points: 150 for MSP and 130 for Max Logit. For each model-classifier pair, we computed an unweighted average across 10 data points (five datasets $\times$ two prompt phrasings) to obtain the values in Table 1.

Statistical significance. To determine whether our AUROC results were statistically significant, we used the Mann-Whitney U test (MWU) (Mann & Whitney, 1947; Wilcoxon, 1945). The MWU directly tests the null hypothesis that a classifier's true AUROC is 50% (i.e., random guessing). For us, a significant MWU affirms the hypothesis that our classifier can distinguish between (1) questions where the LLM answered correctly and (2) questions where the LLM answered incorrectly.

For each available combination of model, dataset, prompt phrasing, and classifier, we tested the null hypothesis that the true AUROC was equal to 50%. This resulted in 280 MWUs. Table 1 reports the number of p-values which were below 10^{-4} for each model-classifier pair. The threshold of $\alpha = 10^{-4}$ accounts for a Bonferroni correction (Bonferroni, 1936), which is applied when performing multiple hypothesis tests (in our case, 280) to ensure that the chance of falsely rejecting *any* null hypothesis is small. Starting from the standard threshold of $\alpha = 0.05$, the Bonferroni correction yields $\alpha = 0.05/280 \approx 1.8 \times 10^{-4}$. We use the stricter threshold of 10^{-4} for simplicity.[12]

5.2 Results

Among these 280 data points, AUROC outperformed random chance with $p < 10^{-4}$ in 233 cases. When the Falcon and Llama 2 models are excluded, that statistic improves to 197/200. These results demonstrate that the MSP and Max Logit are statistically predictive of correctness (except for possibly the weakest models). However, this is perhaps not so surprising given the monotonicity of the calibration curves in Figure 2.

In our opinion, the more exciting finding is a strong direct correlation between average Q&A accuracy and average AUROC: the PCCs for MSP AUROC and Max Logit AUROC were $r = 0.81$ and $r = 0.87$, both with $p < 10^{-3}$ (Figure 3). This finding suggests that as chat LLMs become more powerful, the uncertainty information present in MSPs actually becomes more refined. In contrast, we found no evidence that calibration error will similarly improve (Figure 2, right).

An outlier dataset: WinoGrande. WinoGrande was by far the hardest dataset for our correctness prediction task (Table 2). Our best hypothesis for this discrepancy is that WinoGrande is intentionally adversarial and tries to "trick" the model. An illustrative question from this dataset is "Neil told Craig that he has to take care of the child for the day because _ did it last time." Even for some humans, it could be unclear whether Neil is assuming responsibility or assigning responsibility. One wrinkle is that the Q&A accuracy on WinoGrande is comparable to other datasets, so it is not the case that this dataset is "harder" in general: it is harder only for predicting correctness. Regardless, despite WinoGrande's average MSP AUROC of 57.9%, Llama 3.1 70B and GPT-4o Turbo still achieved AUROCs of 75.8% and 78.3% respectively on this dataset (Table 12), suggesting that this difficulty is surmountable for capable models.

Minimal correlation between model size and AUROC. The PCCs between model size and AUROC were $r = 0.57$ ($p = 0.04$) and $r = 0.36$ ($p = 0.23$) for MSP and Max Logit, respectively (Figure 4). It is unsurprising that some correlation exists, due to the known correlation between size and model capabilities (e.g., Q&A accuracy) and our correlation between Q&A accuracy and AUROC. However, the relatively weak correlation between model size and AUROC may suggest that adding more parameters does not directly improve predictive power of these classifiers outside of improving the model's overall.

Prompt phrasing had minimal impact. The two different prompt phrasings (Figures 1 and 6) did not yield significantly different results (Figure 5). This suggests that our results are robust to minor modifications to prompt phrasing.

[12] We handle the cross-model comparisons separately since we test only 8 cross-model hypotheses (Figure 2, Figure 3, and 6 more in the appendix). As such, we use $\alpha = 0.05/8 \approx 0.006$ for those comparisons. This is not a crucial point since the p-values in Figures 2 and 3 are either above 0.3 or below 10^{-3}.

Table 1: Main AUROC results. AUROC and Q&A values are percentages, averaged over ten data points (five datasets and two phrasings). The $p < 10^{-4}$ columns indicate how many of those ten data points yielded p-values below 10^{-4} for the null hypothesis that AUROC = 50%. The p-values are from the Mann-Whitney U test; see Section 5 for details.

		MSP		Max Logit	
LLM	Q&A Accuracy	AUROC	$p < 10^{-4}$	AUROC	$p < 10^{-4}$
Falcon 7B	29.4	52.5	2/10	51.3	0/10
Falcon 40B	43.8	59.3	7/10	54.8	6/10
Llama 2 7B	40.4	56.9	6/10	55.7	5/10
Llama 2 70B	58.1	69.3	10/10	63.7	8/10
Llama 3.0 8B	58.5	71.3	10/10	68.9	10/10
Llama 3.0 70B	78.4	81.7	10/10	72.6	10/10
Llama 3.1 8B	62.8	72.7	10/10	67.1	10/10
Llama 3.1 70B	80.8	83.3	10/10	67.4	10/10
Mistral 7B	56.3	64.2	10/10	63.5	10/10
Mixtral 8x7B	69.0	60.2	9/10	61.2	10/10
SOLAR 10.7B	67.1	59.7	9/10	65.5	10/10
Yi 6B	53.0	66.6	10/10	60.5	10/10
Yi 34B	70.1	63.8	9/10	64.6	10/10
GPT-3.5 Turbo	67.5	75.6	10/10	–	–
GPT-4o	86.6	84.2	10/10	–	–

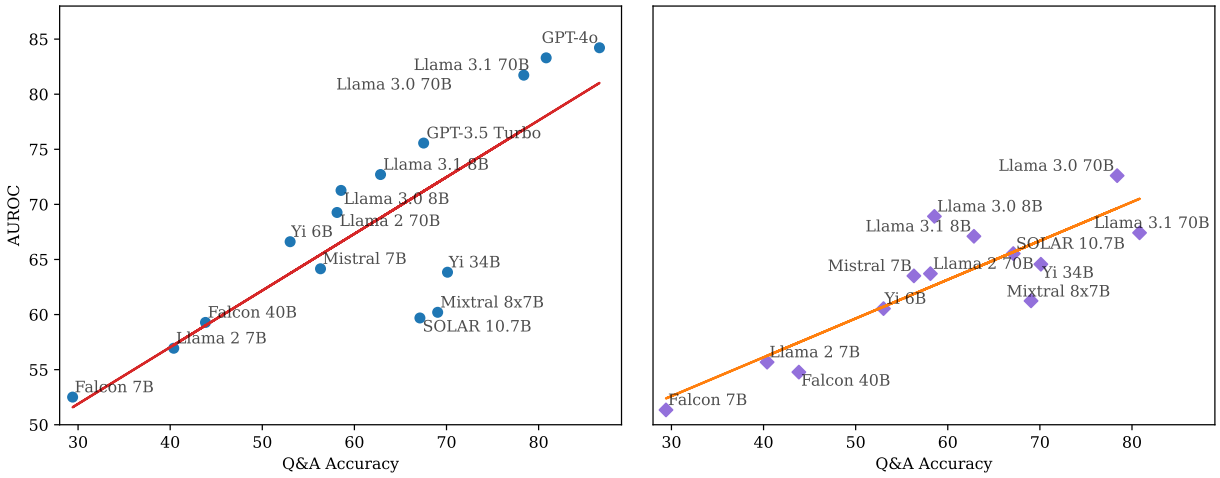

Figure 3: Average AUROC vs average Q&A accuracy for the MSP (left) and Max Logit (right). These plots use the same data as Table 1. The PCCs for MSP and Max Logit were $r = 0.81$ and $r = 0.87$ respectively, both with $p < 10^{-3}$, indicating strong correlations.

Table 2: Q&A accuracy and AUROCs per dataset. All values are percentages, averaged over all models and prompt phrasings.

	Q&A accuracy	MSP AUROC	Max Logit AUROC
ARC-Challenge	69.5	71.9	67.2
HellaSwag	58.3	67.8	62.2
MMLU	54.4	68.5	64.1
TruthfulQA	45.8	66.6	62.0
WinoGrande	59.3	57.9	54.5

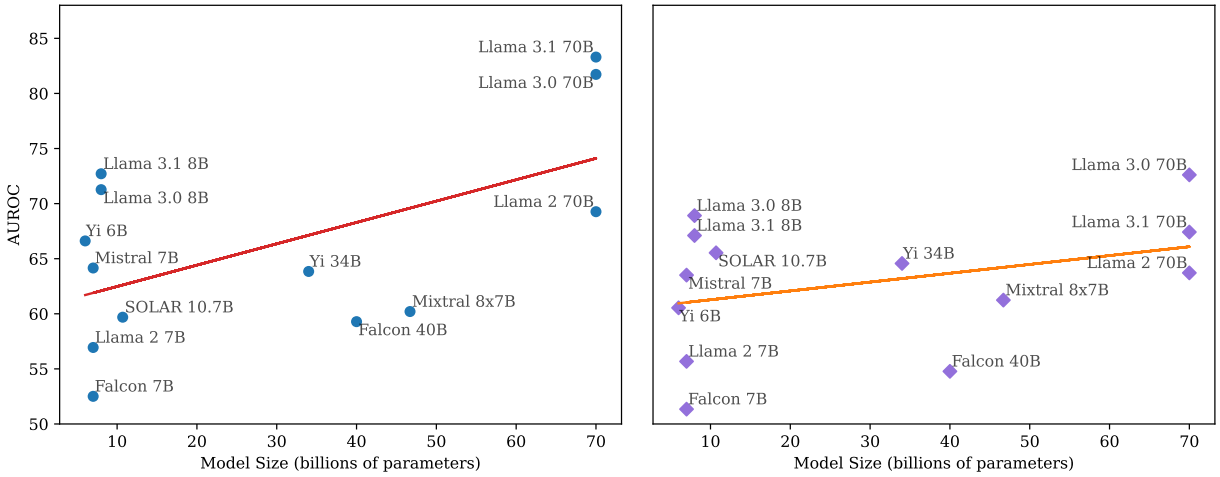

Figure 4: AUROC vs model size for MSP (left) and Max Logit (right). The PCCs for MSP and Max Logit were $r = 0.57$ ($p = 0.04$) and $r = 0.36$ ($p = 0.23$) respectively. GPT-3.5 Turbo and GPT-4o were excluded since their sizes are unknown.

6 Proof-of-concept: reducing wrong answers by abstention

In Section 5, we showed that the MSP and maximum logit can predict correctness. To illustrate the utility of this finding, we now revisit the original Q&A task but allow models to selectively abstain based on the MSP or Max Logit.

6.1 Methodology

These experiments use the same data as the AUROC experiments, but the data is analyzed differently. For each classifier (MSP or Max Logit) and a given threshold, we conducted the following analysis. First, we computed the classifier value (MSP or maximum logit) based on the initial LLM response, the same way we did in our AUROC experiments. If the classifier value was below the threshold, we recorded the model's answer as "abstain" and otherwise recorded the original answer. We awarded 1 point per correct answer, 0 points per abstention, and $-c$ points per wrong answer, normalized by the total number of questions. We performed this analysis for $c = 1$ ("balanced score") and $c = 2$ ("conservative score"). For $c = 1$, the benefit of a correct answer equals the cost of a wrong answer. However, wrong answers are often much worse (e.g., medical diagnoses), justifying $c = 2$. Our experiment design was partly inspired by Kang et al. (2024), who use an even more extreme penalty of $c = 4$.

Choosing the threshold. Unlike our AUROC results, here we must choose a specific threshold for whether the model should abstain. To do so, we randomly selected a training set of k questions, and then for each model, chose the threshold which performed best on those k questions. We discovered that $k = 20$ performed

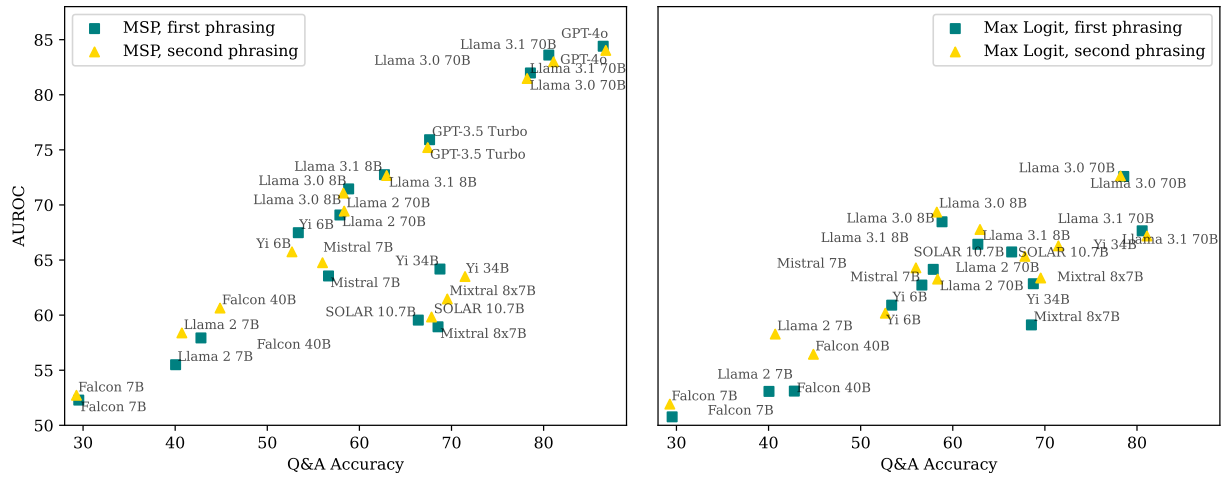

Figure 5: Average AUROC vs Q&A accuracy based on prompt phrasing (the two phrasings can be found in Figures 1 and 6). All values are averaged over the five datasets.

almost as well as using half of all questions. Although $k = 10$ still outperformed the base LLMs, performance was much worse than $k = 20$. For this reason, we used $k = 20$ in our main experiments. We also report results for $k = 10$ (Table 6) and $k =$ half of all questions (Table 7) in the appendix. All tables and figures other than Tables 6 and 7 use $k = 20$.

6.2 Results

For each combination of LLM, classifier, and $c \in \{1, 2\}$, Table 3 reports the scores obtained by the base LLM and by our method on the test set, where our method used the threshold determined by the training set. Figure 8 shows each model's score across the entire range of possible thresholds. Our method outperformed or matched the base LLM in nearly all conditions and substantially outperformed the base LLM on the conservative score metric. (The sole exception is the Llama 3.1 8B, Max Logit, on the balanced score metric, which exhibits a tiny drop in score of 0.3).

As expected, models with low initial scores exhibited the most dramatic improvements. For example, any model with a negative initial score can trivially improve to 0 by abstaining on every question. More generally, the higher the fraction of correct answers, the more likely the model is to accidentally abstain on a correct answer. As a result, it is unsurprising that models with high initial scores showed more modest improvements and lower abstention rates.

Abstention frequency. In line with the reasoning above, some models do abstain quite frequently, with some of the weakest models reaching nearly 100% abstention rates (Table 5 in the appendix). One may be concerned that excessively frequent abstention could render a model unusable. However, we argue that excessively frequent wrong answers would render a system not only unusable but actively harmful. If a model is likely to wrong more often than not, perhaps it is appropriate for the model to always abstain.

Overall, our Q&A-with-abstention results show how the uncertainty signals from softmax probabilities and/or logits can be leveraged to improve performance on practical language tasks. Further details on these results can be found in Appendix A.2.

7 Conclusion

In this paper, we showed that the MSPs of chat LLMs are miscalibrated but still provide a reliable signal of uncertainty. Furthermore, the reliability of this signal improves as model capabilities improve, but the same is not true of calibration.

Table 3: Results on Q&A with abstention. "Balanced" and "conservative" correspond to -1 and -2 points per wrong answer, respectively. Correct answers and abstentions are always worth +1 and 0 points, respectively. The total number of points is divided by the total number of questions (then scaled up by 100 for readability) to obtain the values shown in the table. We highlight the best method(s) for each model.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−41.2	**−0.5**	**−0.5**	−111.8	−5.7	**−1.6**
Falcon 40B	−12.3	**1.2**	0.2	−68.5	−1.5	**−0.5**
Llama 2 7B	−19.4	−11.1	**−8.5**	−79.0	−1.5	**−0.1**
Llama 2 70B	16.2	**21.3**	19.1	−25.7	2.8	**3.0**
Llama 3.0 8B	17.0	**23.5**	20.7	−24.4	**10.3**	10.0
Llama 3.0 70B	**56.8**	**56.8**	**56.8**	35.2	**46.5**	42.2
Llama 3.1 8B	25.6	**29.6**	25.3	−11.5	**18.0**	13.2
Llama 3.1 70B	61.6	**62.8**	61.6	42.5	**52.0**	41.1
Mistral 7B	12.7	**16.7**	15.1	−31.0	−5.0	**5.0**
Mixtral 8x7B	38.1	**38.7**	34.2	7.2	12.7	**14.8**
SOLAR 10.7B	**34.2**	**34.2**	**34.2**	1.3	4.7	**10.6**
Yi 6B	6.0	**15.6**	9.7	−41.0	**6.5**	1.1
Yi 34B	40.1	**41.1**	40.2	10.2	**14.2**	11.9
GPT-3.5 Turbo	35.0	**38.0**	–	2.5	**27.9**	–
GPT-4o	73.2	**73.5**	–	59.9	**61.4**	–

However, our study has several important limitations. One is the restriction to multiple-choice questions, which simplifies the problem in several ways. First, it removes the need to distinguish between multiple correct answers ("aleatoric uncertainty") and no good answers ("epistemic uncertainty"), both of which could result in low MSPs. The restriction to multiple-choice also enabled us to tie the LLM's answer to a single token and thus a single MSP. Future work could handle free-response questions by aggregating MSPs across tokens in a clever way. A further challenge could be multi-step decision-making problems which may involve aggregating uncertainty not only across multiple tokens in a single response, but across multiple responses on different time steps.

Another limitation is our reliance on labeled data to transform our scientific insights into a practical method for abstention. We only used 20 data points, and labeled data was only used to choose the threshold, but a fully unsupervised method would be advantageous in many settings. However, we remind the reader that our primary contribution is the scientific finding of good correctness prediction despite miscalibration, not the proof-of-concept abstention experiments in Section 6.

Finally, we would also like to better understand when and why these methods fail. Are there particular subcategories of unfamiliar situations that are especially challenging to identify? For example, why was the WinoGrande dataset so much harder for our correctness prediction task?

We think that these limitations correspond to exciting avenues for future work. More broadly, we are excited about developing more robust methods for mistake detection in LLMs, both for Q&A tasks and for other contexts.

Broader impact statement

The capabilities of AI systems have advanced rapidly over the past several years and will likely continue to grow. In order to ensure that AI is beneficial for society, we believe it is paramount to understand the risks of such systems and take steps to address those risks. In this paper, we focus on one particular risk:

harmful responses from LLMs. We hope that our work contributes the ongoing efforts to mitigate harmful LLM responses.

We do not think it is likely for our work to inadvertently cause harm, but one possibility is worth mentioning. If a reader were to assume that abstention methods can completely eliminate false responses, that reader might be more likely to fall prey to false responses when they do inevitably still occur. We caution readers to remain vigilant about false responses from LLMs.

References

01-ai. 01-ai/Yi-6B-Chat · Hugging Face, 2023. URL https://huggingface.co/01-ai/Yi-6B-Chat. Accessed Jan 2024.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open LLM Leaderboard. https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard, 2023.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(96)00142-2. URL https://www.sciencedirect.com/science/article/pii/S0031320396001422.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, January 1970. ISSN 1557-9654. doi: 10.1109/TIT.1970.1054406. URL https://ieeexplore.ieee.org/abstract/document/1054406. Conference Name: IEEE Transactions on Information Theory.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *NAACL-HLT*, 2024.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language Model Cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations*, ICLR '24, April 2024. doi: 10.48550/arXiv.2404.10136. URL http://arxiv.org/abs/2404.10136. arXiv:2404.10136 [cs].

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR '17)*, October 2017. URL http://arxiv.org/abs/1610.02136.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR '21)*, 2021.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8759–8773. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html. ISSN: 2640-3498.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

S Kadavath, T Conerly, A Askell, T Henighan, D Drain, E Perez, N Schiefer, ZH Dodds, N DasSarma, E Tran-Johnson, et al. Language models (mostly) know what they know. *URL https://arxiv. org/abs/2207.05221*, 5, 2022.

Katie Kang, Amrith Setlur, Claire Tomlin, and Sergey Levine. Deep Neural Networks Tend To Extrapolate Predictably. In *The Twelfth International Conference on Learning Representations*, ICLR '24, 2024.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling, December 2023. URL http://arxiv.org/abs/2312.15166. arXiv:2312.15166 [cs].

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Dan Mangan. Judge sanctions lawyers for brief written by A.I. with fake citations. *CNBC*, June 2023. URL www.cnbc.com/2023/06/22/judge-sanctions-lawyers-whose-ai-written-filing-contained-fake-citations.html. Accessed Jan 2024.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-01-15.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: an adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://dl.acm.org/doi/10.1145/3474381.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*, 2023.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, 2023.

```
You will be asked a multiple-choice question.  Respond with the letter which corresponds
to the correct answer, followed by a period.  There is no need to provide an explanation,
so your response should be very short.  Now here is the question:

In the nitrogen cycle, nitrogen can return to the lithosphere directly from the atmosphere
by
A. lightning.
B. cellular respiration.
C. air pollution.
D. condensation.

Answer:
```

Figure 6: The second prompt phrasing we used.

Table 4: Q&A accuracy and calibration error (scaled by 100 for readability) for each model.

LLM	Q&A accuracy	Calibration error ($\times 100$)
Falcon 7B	29.4	7.8
Falcon 40B	43.8	15.1
Llama 2 7B	40.4	32.2
Llama 2 70B	58.1	32.2
Llama 3.0 8B	58.5	22.5
Llama 3.0 70B	78.4	22.4
Llama 3.1 8B	62.8	12.5
Llama 3.1 70B	80.8	9.2
Mistral 7B	56.3	38.4
Mixtral 8x7B	69.0	29.8
SOLAR 10.7B	67.1	25.5
Yi 6B	53.0	19.5
Yi 34B	70.1	11.5
GPT-3.5 Turbo	67.5	22.1
GPT-4o	86.6	16.7

A Details on main experiments

Here we include details on the methodology and results of our main experiments. First, we include the second prompt phrasing we used (Figure 6). Recall that Figure 1 shows the first phrasing.

A.1 Details on calibration results

As mentioned in Section 4, each model's calibration error was computed by first computing the calibration error for each combination of model-dataset-phrasing, and then averaging over datasets and phrasings to obtain a per-model total error. This was done to avoid downweighting datasets with fewer questions. This data is shown in Figure 2 (right) in Section 4 and also in Table 4 here in this appendix.

However, this approach does not make sense for the calibration curves in Figure 2. This is because calibration curves are not averages: they are obtained by bucketing each data point (in our case, a question-response pair is a data point). In order to avoid downweighting the smaller datasets, we duplicated data points from those datasets so that the total number of points per dataset was roughly 6000. Concretely, we duplicated

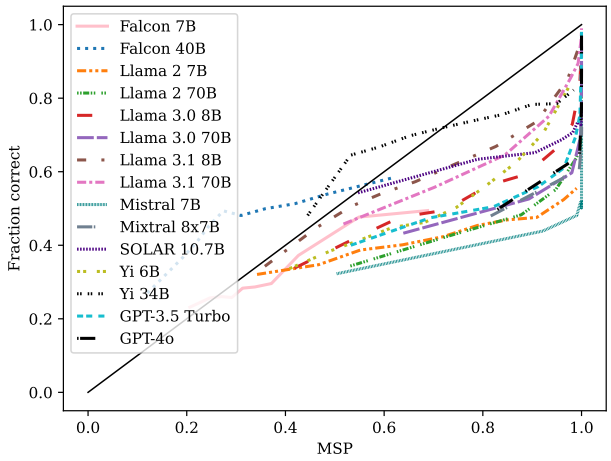

Figure 7: A variant of Figure 2 where smaller datasets are not duplicated and thus are downweighted.

data points from TruthfulQA by $6000/817 \approx 7$ and from ARC by $6000/2590 \approx 2$. Although this solution has drawbacks, we felt that it was the least bad option. The data in Figure 2 uses this duplication approach.

For the curious reader, Figure 7 provides version where the data is not duplicated (and as a result, down-weights TruthfulQA and ARC). Qualitatively, there are no significant differences between this figure and the duplication approach in Figure 2.

A.2 Details on Q&A-with-abstention results

Figure 8 is based on the same data as Table 3, but shows each model's performance across the entire range of possible thresholds. A threshold of zero corresponds to the base LLM and the black dot indicates the threshold chosen during the training phase (using 20 labeled data points), which is also the threshold used to compute the score in Table 3 using 20 data points. One can see that the chosen thresholds are not quite optimal, but 20 data points was still enough to produce substantial improvements over the baseline of not abstaining.

Table 5 presents the average abstention frequencies across all datasets, corresponding to the scores from Table 3. We also provide dataset-level versions of Table 5 in Appendix F.

Finally, Tables 6 and 7 are analogues of Table 3 using 10 data points and half of all data points for training, respectively.

B Results for 5-shot prompting

As discussed in Section 3, we chose zero-shot prompting for our main results in order to test our hypothesis in the simplest setting possible. In this section, we show how the results change if 5-shot prompting is used instead. The short answer is that the same correlations and trends hold, except for Max Logit. The AUROC values are also slightly lower overall. This could suggest that few-shot prompting partially disrupts the innate uncertainty information in LLMs, but we do not have the evidence to conclude this definitively.

Figure 9 presents the calibration results for 5-shot prompting. As in the zero-shot setting, models are consistently overconfident and there is no correlation between calibration error and Q&A accuracy ($r = -0.02, p = 0.95$). If anything, calibration error is slightly higher in the 5-shot setting, but this may just be noise.

Figure 10 presents the AUROC correctness prediction results for 5-shot prompting. The correlation between MSP AUROC and Q&A accuracy remains ($r = 0.69, p < 0.005$), but Max Logit AUROC no longer exhibits a statistically significant correlation with Q&A accuracy ($r = 0.43, p = 0.14$).

Table 5: Frequency of abstention in the Section 6 experiments.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0	89.9	98.4	0	89.9	98.4
Falcon 40B	0	90.4	98.4	0	94.4	98.4
Llama 2 7B	0	21.6	32.7	0	98.0	99.1
Llama 2 70B	0	15.4	11.9	0	36.4	64.2
Llama 3.0 8B	0	31.2	58.1	0	52.0	58.1
Llama 3.0 70B	0	0.0	0.0	0	42.4	26.5
Llama 3.1 8B	0	40.8	33.9	0	58.2	65.6
Llama 3.1 70B	0	15.6	0.0	0	15.6	17.7
Mistral 7B	0	38.1	8.2	0	41.8	74.2
Mixtral 8x7B	0	9.3	27.1	0	9.3	27.1
SOLAR 10.7B	0	0.0	0.0	0	5.8	15.2
Yi 6B	0	52.7	44.1	0	77.0	98.1
Yi 34B	0	5.0	3.3	0	5.0	3.3
GPT-3.5 Turbo	0	15.3		0	45.6	
GPT-4o	0	2.2		0	2.2	

Table 6: Q&A with abstention results for $k = 10$. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−41.2	−0.4	−1.0	−111.8	−5.7	−1.6
Falcon 40B	−12.4	−7.4	−8.2	−68.6	−14.7	−16.8
Llama 2 7B	−19.3	−11.0	−17.8	−79.0	−3.0	−8.7
Llama 2 70B	16.2	21.3	19.1	−25.7	7.4	3.0
Llama 3.0 8B	17.1	22.6	17.2	−24.4	10.4	8.3
Llama 3.0 70B	56.8	56.8	56.8	35.2	47.3	35.2
Llama 3.1 8B	25.7	25.8	26.0	−11.5	3.8	−8.8
Llama 3.1 70B	61.7	62.9	61.7	42.5	51.5	42.5
Mistral 7B	12.6	16.9	15.5	−31.0	−6.8	−4.2
Mixtral 8x7B	38.1	38.8	35.4	7.2	12.6	14.9
SOLAR 10.7B	34.2	34.2	34.2	1.3	1.3	9.9
Yi 6B	6.0	15.5	10.4	−41.0	6.5	1.1
Yi 34B	40.2	40.3	40.3	10.2	14.3	10.8
GPT-3.5 Turbo	35.0	38.5		2.5	27.3	
GPT-4o	73.2	73.3		59.9	65.8	

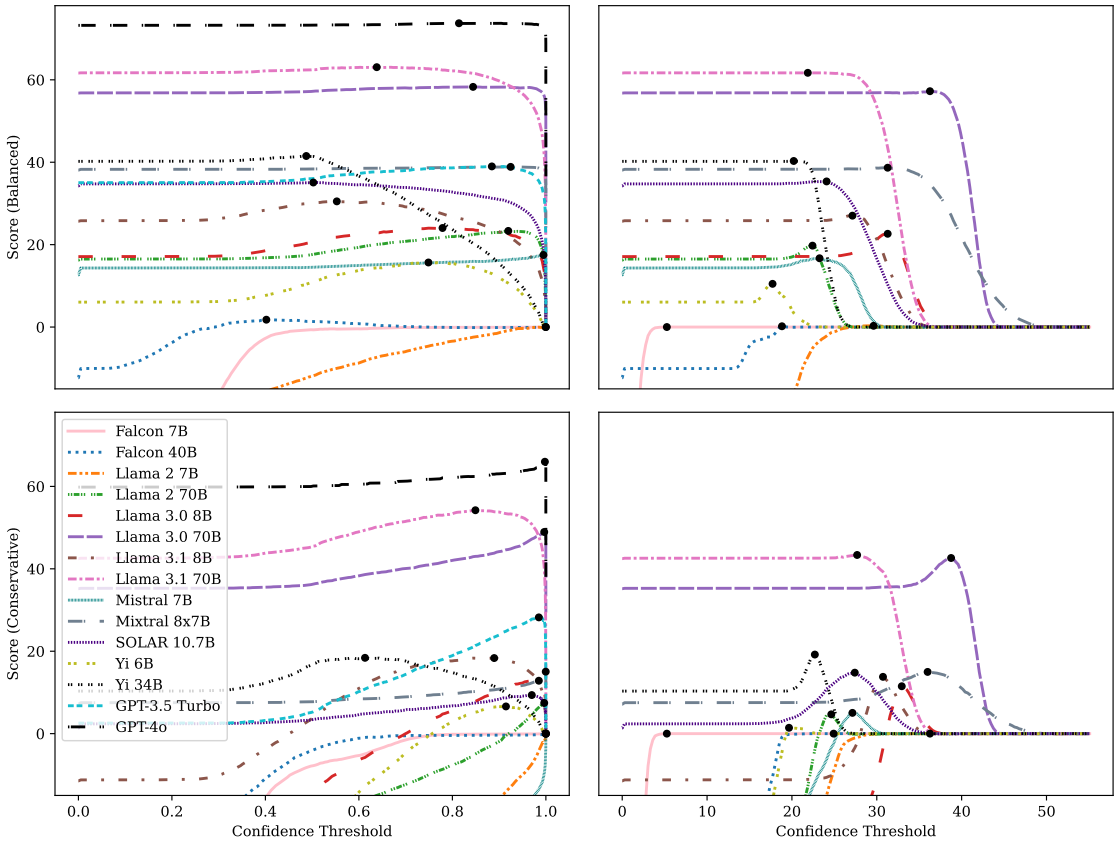

Figure 8: Q&A-with-abstention scores across all possible thresholds. The base LLM corresponds to a threshold of zero. The black dot indicates the threshold selected via the training set, which determines the MSP and Max Logit scores in Table 3.

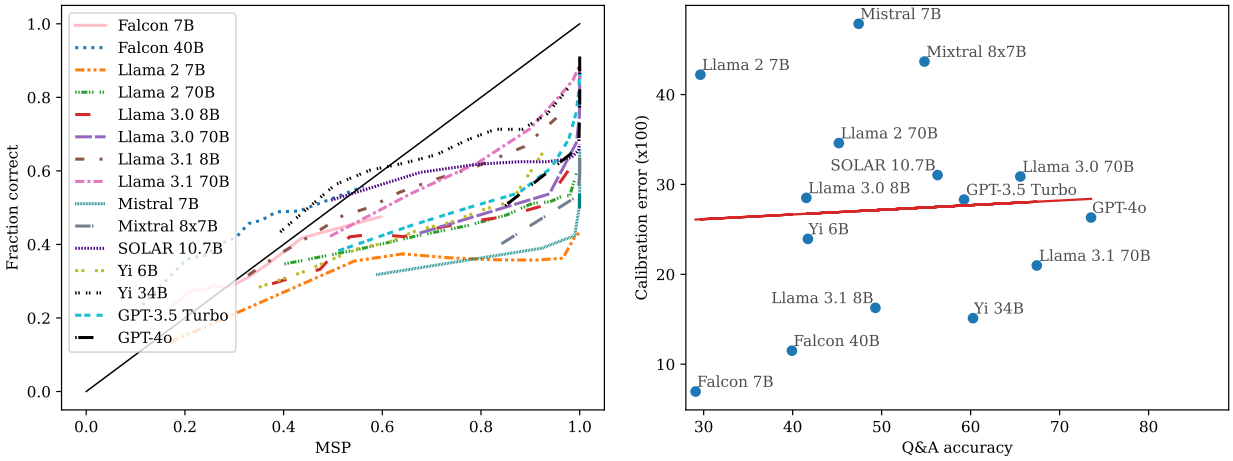

Figure 9: Calibration results for 5-shot prompting. Left: calibration curves for the 15 models. Right: as for zero-shot, there is no correlation between calibration error and average Q&A accuracy ($r = -0.02, p = 0.95$).

Table 7: Q&A with abstention results for k = half of all data points. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−41.0	0.0	0.0	−111.5	−0.1	0.0
Falcon 40B	−12.4	1.9	0.2	−68.5	0.0	0.0
Llama 2 7B	−19.1	0.0	0.3	−78.7	0.0	−0.1
Llama 2 70B	16.9	23.5	20.2	−24.7	7.3	4.8
Llama 3.0 8B	17.1	23.9	23.3	−24.4	12.6	11.2
Llama 3.0 70B	56.1	57.8	56.5	34.2	48.4	41.9
Llama 3.1 8B	25.5	30.3	26.9	−11.7	17.7	13.1
Llama 3.1 70B	61.4	62.8	61.4	42.1	53.8	42.6
Mistral 7B	11.4	16.2	15.3	−32.8	0.0	4.9
Mixtral 8x7B	37.4	38.4	37.8	6.0	13.5	13.6
SOLAR 10.7B	33.7	34.5	34.5	0.6	8.4	13.5
Yi 6B	5.0	14.6	9.7	−42.4	6.1	1.1
Yi 34B	40.4	41.6	40.4	10.6	18.5	19.1
GPT-3.5 Turbo	34.8	38.9		2.1	28.4	
GPT-4o	73.4	73.9		60.1	66.1	

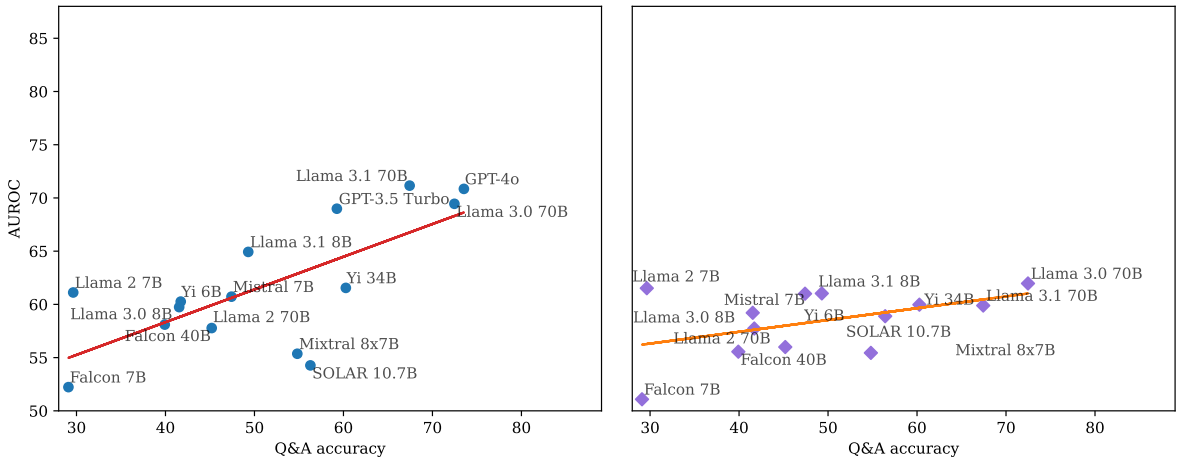

Figure 10: AUROC results for 5-shot prompting. MSP AUROC (left) still exhibits a strong correlation with Q&A accuracy ($r = 0.62, p < 0.005$). However, Max Logit AUROC (right) no longer exhibits a statistically significant correlation with average Q&A accuracy ($r = 0.43, p = 0.14$).

C What if we use the product of MSPs across tokens?

As discussed in Section 3, an LLM response typically consists of multiple tokens, each with its own MSP. For our main results, we selected the MSP of the token which indicates the LLM's answer, i.e., "A"/"B"/etc. We chose this process because we believe that it is the fairest way of computing the probability that the model assigns to its answer. However, it is important to know whether our results would change if the MSP were computed differently. In this section, we consider one potential alternative: computing the product of

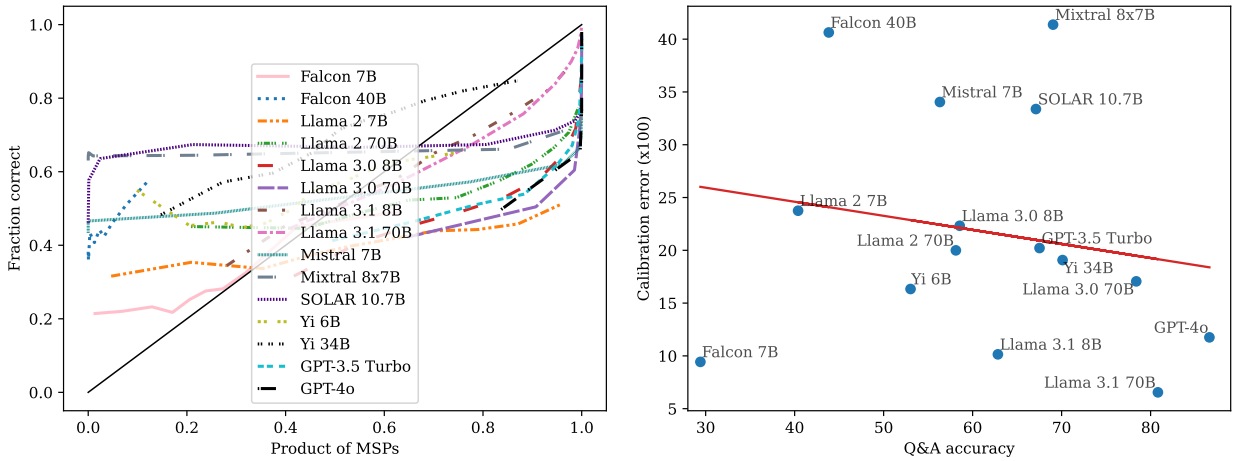

Figure 11: Calibration results for the product of MSPs. There is still no evidence for a correlation between calibration error and Q&A accuracy ($r = -0.19, p = 0.50$).

MSPs across all tokens. This value corresponds to the overall probability the model assigns to its output.[13] Does this method produce different results?

The answer is generally no, for both calibration (Figure 11) and correctness prediction (Figure 12). The biggest difference is that the product of MSPs is typically smaller than the MSP of the answer token, which is expected. As a result, the LLMs are no longer consistently overconfident. However, the models remain consistently miscalibrated: they are now simply underconfident in addition to being overconfident. (The underconfident instances may correspond to responses where the LLM included an explanation of its answer, which would cause the product of MSPs to be artificially low.) Also, there remains no correlation between calibration error and Q&A accuracy ($r = -0.19, p = 0.50$).

With regards to correctness prediction, we find the same correlation between AUROC and Q&A accuracy with approximately the same strength ($r = 0.83, p = 0.00014$).

These results show that our findings are robust to modifications in how the MSP is computed across tokens.

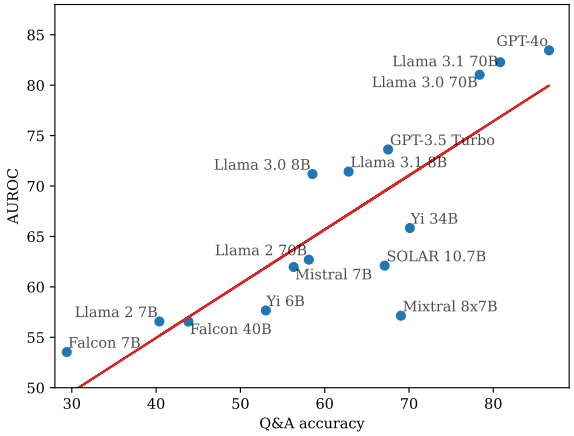

Figure 12: AUROC results for the product of MSPs. There remains a strong correlation between AUROC and Q&A accuracy ($r = 0.83, p < 10^{-3}$).

[13]Since the product of maximum logits does not admit such an interpretation, we only study the product of MSPs here.

D Licenses

The Hugging Face transformers library uses the Apache 2.0 license. The licenses for the datasets we used are:

1. ARC-Challenge: CC BY-SA 4.0
2. HellaSwag: MIT
3. MMLU: MIT
4. TruthfulQA: Apache 2.0
5. WinoGrande: CC-BY

The licenses for the models we used are:

1. Falcon 7B and 40B: Apache 2.0
2. Llama 2 7B and 70B: Llama 2 Community License Agreement
3. Llama 3 8B and 70B: Llama 3 Community License Agreement
4. Llama 3.1 8B and 70B: Llama 3.1 Community License Agreement
5. Mistral 7B and Mixtral 8x7B: Apache 2.0
6. SOLAR 10.7B: CC BY-NC 4.0
7. Yi 6B and 34B: Yi Series Models Community License Agreement
8. GPT-3.5 Turbo and GPT-4o: OpenAI Terms of Service

All models and datasets were used in a manner consistent with their licenses and intended use.

E Caveat for Falcon 7B

Initially, many of Falcon 7B's responses fell into the "unparseable" category described in Section 3. Upon investigation, we found that many of these responses were simply a period or an end-of-text token. Removing the final newline in the prompt resolved this behavior, so we believe that this newline was somehow convincing the model that the conversation was "over". These initial results had the side effect of making it very easy to detect wrong answers, since a solitary period is obviously not a correct answer. For this reason, we removed the final newline for Falcon 7B only. We considered removing the final newline for all models or excluding Falcon 7B entirely, but we felt that our chosen approach would be more scientifically honest. As Falcon 7B performed by far the worst on both Q&A and AUROC even with this concession, we do not think this decision holds much import, but we report it for transparency.

F Dataset-level results

For the curious reader, here we present dataset-level versions of Table 1 (AUROC and Q&A accuracy), Table 3 (results for Q&A with abstention), and Table 5 (frequency of abstention on Q&A with abstention experiments).

Table 8: AUROC results for ARC-Challenge. See Table 1 for more explanation.

		MSP		Max Logit	
LLM	Q&A Accuracy	AUROC	$p < 10^{-4}$	AUROC	$p < 10^{-4}$
Falcon 7B	25.7	50.5	0/2	50.2	0/2
Falcon 40B	53.0	66.8	2/2	59.8	2/2
Llama 2 7B	45.2	62.8	2/2	59.8	2/2
Llama 2 70B	74.2	77.4	2/2	69.5	2/2
Llama 3.0 8B	74.7	81.8	2/2	78.7	2/2
Llama 3.0 70B	92.3	89.7	2/2	82.2	2/2
Llama 3.1 8B	77.4	82.2	2/2	77.6	2/2
Llama 3.1 70B	93.4	91.3	2/2	78.6	2/2
Mistral 7B	70.7	67.3	2/2	67.3	2/2
Mixtral 8x7B	84.2	63.5	2/2	64.3	2/2
SOLAR 10.7B	80.5	61.4	2/2	71.2	2/2
Yi 6B	68.8	75.6	2/2	64.1	2/2
Yi 34B	85.4	67.3	2/2	70.0	2/2
GPT-3.5 Turbo	83.3	84.4	2/2		2/2
GPT-4o	96.3	88.4	2/2		2/2

Table 9: AUROC results for HellaSwag. See Table 1 for more explanation.

		MSP		Max Logit	
LLM	Q&A Accuracy	AUROC	$p < 10^{-4}$	AUROC	$p < 10^{-4}$
Falcon 7B	24.9	50.1	0/2	49.4	0/2
Falcon 40B	42.8	61.8	2/2	59.2	2/2
Llama 2 7B	41.5	57.3	2/2	53.8	1/2
Llama 2 70B	63.9	70.0	2/2	64.7	2/2
Llama 3.0 8B	65.6	73.5	2/2	70.8	2/2
Llama 3.0 70B	77.6	81.6	2/2	70.4	2/2
Llama 3.1 8B	66.1	72.9	2/2	67.5	2/2
Llama 3.1 70B	78.9	81.7	2/2	59.8	2/2
Mistral 7B	52.5	63.0	2/2	62.6	2/2
Mixtral 8x7B	65.8	57.9	2/2	59.0	2/2
SOLAR 10.7B	79.2	63.5	2/2	69.0	2/2
Yi 6B	42.0	67.0	2/2	60.6	2/2
Yi 34B	75.4	71.8	2/2	68.7	2/2
GPT-3.5 Turbo	72.7	77.0	2/2		2/2
GPT-4o	88.4	87.9	2/2		2/2

Table 10: AUROC results for MMLU. See Table 1 for more explanation.

		MSP		Max Logit	
LLM	Q&A Accuracy	AUROC	$p < 10^{-4}$	AUROC	$p < 10^{-4}$
Falcon 7B	25.8	50.1	0/2	50.9	0/2
Falcon 40B	43.9	62.0	2/2	52.9	1/2
Llama 2 7B	40.2	61.7	2/2	59.9	2/2
Llama 2 70B	56.0	71.8	2/2	65.9	2/2
Llama 3.0 8B	56.3	76.2	2/2	74.5	2/2
Llama 3.0 70B	75.0	83.5	2/2	78.6	2/2
Llama 3.1 8B	58.1	78.1	2/2	73.9	2/2
Llama 3.1 70B	79.8	84.1	2/2	72.9	2/2
Mistral 7B	52.5	64.9	2/2	64.8	2/2
Mixtral 8x7B	65.2	60.6	2/2	64.0	2/2
SOLAR 10.7B	58.0	60.8	2/2	66.9	2/2
Yi 6B	51.9	69.8	2/2	62.3	2/2
Yi 34B	65.6	64.1	2/2	65.7	2/2
GPT-3.5 Turbo	65.4	79.8	2/2		2/2
GPT-4o	84.0	85.2	2/2		2/2

Table 11: AUROC results for TruthfulQA. See Table 1 for more explanation.

		MSP		Max Logit	
LLM	Q&A Accuracy	AUROC	$p < 10^{-4}$	AUROC	$p < 10^{-4}$
Falcon 7B	20.5	62.4	2/2	55.8	0/2
Falcon 40B	28.7	53.4	0/2	50.6	0/2
Llama 2 7B	24.1	52.2	0/2	53.9	0/2
Llama 2 70B	44.6	71.5	2/2	66.9	2/2
Llama 3.0 8B	40.4	68.0	2/2	64.0	2/2
Llama 3.0 70B	72.1	79.8	2/2	70.9	2/2
Llama 3.1 8B	55.3	72.8	2/2	60.8	2/2
Llama 3.1 70B	74.1	83.6	2/2	66.2	2/2
Mistral 7B	55.1	66.0	2/2	63.4	2/2
Mixtral 8x7B	67.3	65.6	2/2	64.5	2/2
SOLAR 10.7B	49.1	58.1	1/2	63.1	2/2
Yi 6B	43.7	64.5	2/2	60.6	2/2
Yi 34B	56.7	66.8	2/2	67.8	2/2
GPT-3.5 Turbo	55.4	74.8	2/2		2/2
GPT-4o	82.4	81.3	2/2		2/2

Table 12: AUROC results for WinoGrande. See Table 1 for more explanation.

		MSP		Max Logit	
LLM	Q&A Accuracy	AUROC	$p < 10^{-4}$	AUROC	$p < 10^{-4}$
Falcon 7B	50.1	49.4	0/2	50.4	0/2
Falcon 40B	50.8	52.5	1/2	51.5	1/2
Llama 2 7B	51.0	50.8	0/2	50.9	0/2
Llama 2 70B	51.9	55.5	2/2	51.6	0/2
Llama 3.0 8B	55.8	56.8	2/2	56.5	2/2
Llama 3.0 70B	74.9	74.0	2/2	61.0	2/2
Llama 3.1 8B	57.3	57.6	2/2	55.7	2/2
Llama 3.1 70B	77.8	75.8	2/2	59.7	2/2
Mistral 7B	50.8	59.5	2/2	59.5	2/2
Mixtral 8x7B	62.8	53.4	1/2	54.4	2/2
SOLAR 10.7B	68.8	54.7	2/2	57.4	2/2
Yi 6B	58.7	56.1	2/2	55.2	2/2
Yi 34B	67.5	49.2	1/2	50.7	2/2
GPT-3.5 Turbo	60.8	61.8	2/2		2/2
GPT-4o	81.9	78.3	2/2		2/2

Table 13: Q&A with abstention results for ARC-Challenge. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−48.6	−0.9	−0.3	−122.9	−2.2	−0.9
Falcon 40B	5.9	13.7	13.4	−41.1	−9.3	−6.7
Llama 2 7B	−9.6	−5.2	0.2	−64.4	−10.5	−26.2
Llama 2 70B	48.4	48.4	48.4	22.6	39.6	24.6
Llama 3.0 8B	49.3	53.4	50.0	23.9	40.2	34.2
Llama 3.0 70B	84.7	84.7	84.7	77.0	77.0	77.0
Llama 3.1 8B	54.7	57.5	54.6	32.0	45.3	36.8
Llama 3.1 70B	86.8	86.8	86.8	80.2	80.2	80.2
Mistral 7B	41.3	35.4	36.2	12.0	21.8	21.8
Mixtral 8x7B	68.3	68.3	68.3	52.5	52.5	52.5
SOLAR 10.7B	60.9	60.9	62.0	41.4	41.3	46.7
Yi 6B	37.6	41.1	31.2	6.3	26.1	10.9
Yi 34B	70.7	71.2	70.7	56.1	58.3	55.5
GPT-3.5 Turbo	66.7	67.6		50.0	58.7	
GPT-4o	92.6	92.6		88.9	88.9	

Table 14: Frequency of abstention on ARC-Challenge in the Section 6 experiments.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0	98.2	99.2	0	98.2	99.2
Falcon 40B	0	40.1	26.8	0	40.1	85.2
Llama 2 7B	0	10.7	47.1	0	63.7	47.1
Llama 2 70B	0	0.0	0.0	0	28.9	57.0
Llama 3.0 8B	0	20.3	18.4	0	20.3	18.4
Llama 3.0 70B	0	0.0	0.0	0	0.0	0.0
Llama 3.1 8B	0	18.1	9.8	0	18.1	9.8
Llama 3.1 70B	0	0.0	0.0	0	0.0	0.0
Mistral 7B	0	37.2	35.0	0	37.2	35.0
Mixtral 8x7B	0	0.0	0.0	0	0.0	0.0
SOLAR 10.7B	0	0.0	7.5	0	7.1	7.5
Yi 6B	0	18.7	28.1	0	30.8	28.1
Yi 34B	0	2.9	0.0	0	2.9	8.1
GPT-3.5 Turbo	0	6.0		0	23.7	
GPT-4o	0	0.0		0	0.0	

Table 15: Q&A with abstention results for HellaSwag. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−50.2	−11.1	−0.7	−125.3	−4.2	−1.8
Falcon 40B	−14.3	0.8	0.5	−71.4	−0.3	0.2
Llama 2 7B	−17.0	−0.4	−1.1	−75.5	−12.6	−5.0
Llama 2 70B	27.8	32.0	30.3	−8.3	7.7	10.6
Llama 3.0 8B	31.2	35.4	32.0	−3.1	18.3	17.2
Llama 3.0 70B	55.3	55.3	54.5	32.9	38.2	38.4
Llama 3.1 8B	32.2	33.0	32.2	−1.7	20.5	14.4
Llama 3.1 70B	57.9	58.4	57.9	36.8	46.9	36.8
Mistral 7B	5.0	11.1	10.2	−42.5	−7.5	2.0
Mixtral 8x7B	31.6	31.9	26.8	−2.6	4.0	7.1
SOLAR 10.7B	58.4	58.4	58.4	37.6	41.5	37.6
Yi 6B	−15.9	4.5	0.7	−73.9	1.8	−0.1
Yi 34B	50.9	51.1	49.9	26.3	28.5	32.8
GPT-3.5 Turbo	45.4	47.2		18.1	32.9	
GPT-4o	76.9	76.9		65.3	65.3	

Table 16: Frequency of abstention on HellaSwag in the Section 6 experiments.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0	78.2	98.5	0	96.7	98.5
Falcon 40B	0	97.2	98.9	0	97.2	98.9
Llama 2 7B	0	75.9	93.2	0	75.9	93.2
Llama 2 70B	0	19.4	13.9	0	19.4	56.9
Llama 3.0 8B	0	30.6	7.1	0	30.6	55.0
Llama 3.0 70B	0	0.0	13.2	0	6.8	13.2
Llama 3.1 8B	0	1.5	0.0	0	61.6	55.6
Llama 3.1 70B	0	1.5	0.0	0	39.3	0.0
Mistral 7B	0	51.6	64.1	0	51.6	86.5
Mixtral 8x7B	0	12.3	33.8	0	12.3	33.8
SOLAR 10.7B	0	0.0	0.0	0	8.4	0.0
Yi 6B	0	90.1	97.8	0	90.1	97.8
Yi 34B	0	0.7	11.2	0	56.3	28.0
GPT-3.5 Turbo	0	9.6		0	26.3	
GPT-4o	0	0.0		0	0.0	

Table 17: Q&A with abstention results for MMLU. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−48.3	−0.5	0.0	−122.4	−1.3	0.0
Falcon 40B	−12.3	−6.6	−7.6	−68.4	−0.6	−20.6
Llama 2 7B	−19.6	3.0	−13.5	−79.5	−1.4	−0.2
Llama 2 70B	12.1	18.5	13.0	−31.9	−13.2	5.1
Llama 3.0 8B	12.5	12.8	12.5	−31.3	16.4	9.8
Llama 3.0 70B	49.9	53.2	49.9	24.9	43.7	41.3
Llama 3.1 8B	16.3	27.6	19.4	−25.6	21.2	16.8
Llama 3.1 70B	59.6	61.4	59.6	39.5	46.7	42.9
Mistral 7B	5.0	10.8	7.9	−42.6	−24.5	−34.5
Mixtral 8x7B	30.4	31.6	32.9	−4.5	0.8	10.0
SOLAR 10.7B	16.0	16.0	17.5	−26.0	−26.0	−12.2
Yi 6B	3.8	15.4	7.8	−44.3	6.7	−17.5
Yi 34B	31.1	31.1	31.1	−3.4	10.6	−3.4
GPT-3.5 Turbo	30.8	36.0		−3.7	13.7	
GPT-4o	67.9	68.0		51.9	61.9	

Table 18: Frequency of abstention on MMLU in the Section 6 experiments.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0	98.9	100.0	0	98.9	100.0
Falcon 40B	0	11.3	7.7	0	93.3	69.8
Llama 2 7B	0	86.2	13.6	0	95.4	99.0
Llama 2 70B	0	18.2	4.0	0	18.2	68.0
Llama 3.0 8B	0	0.9	0.1	0	55.9	51.8
Llama 3.0 70B	0	9.6	0.0	0	27.2	34.9
Llama 3.1 8B	0	59.8	8.0	0	59.8	74.7
Llama 3.1 70B	0	9.2	0.0	0	9.2	26.6
Mistral 7B	0	18.6	7.4	0	18.6	7.4
Mixtral 8x7B	0	6.9	14.9	0	6.9	22.5
SOLAR 10.7B	0	0.0	4.9	0	0.0	15.4
Yi 6B	0	41.8	41.5	0	65.4	41.5
Yi 34B	0	0.0	0.0	0	35.3	0.0
GPT-3.5 Turbo	0	19.3		0	19.3	
GPT-4o	0	0.7		0	17.1	

Table 19: Q&A with abstention results for TruthfulQA. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	−59.1	−11.3	−2.2	−138.7	−10.3	−5.0
Falcon 40B	−42.6	−3.0	−0.7	−113.9	−7.9	−2.5
Llama 2 7B	−52.4	−39.7	−44.4	−128.6	−4.7	−0.6
Llama 2 70B	−11.2	8.7	4.0	−66.7	−16.8	−19.6
Llama 3.0 8B	−19.5	−5.0	1.4	−79.2	−0.9	−9.1
Llama 3.0 70B	44.2	46.4	39.7	16.4	38.9	28.3
Llama 3.1 8B	10.4	22.2	11.7	−34.4	9.8	−4.6
Llama 3.1 70B	48.2	49.3	48.2	22.3	41.8	26.3
Mistral 7B	10.3	17.6	12.5	−34.6	−8.9	3.3
Mixtral 8x7B	34.7	37.3	32.9	2.0	6.8	14.7
SOLAR 10.7B	−2.0	4.9	8.9	−53.0	−8.6	−8.1
Yi 6B	−12.8	−7.9	−5.9	−69.1	−4.6	−12.1
Yi 34B	13.0	14.5	14.5	−30.5	−0.2	−25.6
GPT-3.5 Turbo	10.4	21.6		−34.4	14.3	
GPT-4o	64.9	65.7		47.4	55.9	

Table 20: Frequency of abstention on TruthfulQA in the Section 6 experiments.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0	72.6	96.8	0	89.1	96.8
Falcon 40B	0	93.4	97.1	0	93.4	97.1
Llama 2 7B	0	21.0	12.3	0	95.8	99.3
Llama 2 70B	0	40.4	34.8	0	40.4	45.6
Llama 3.0 8B	0	28.9	66.0	0	81.6	73.4
Llama 3.0 70B	0	38.6	37.7	0	38.6	37.7
Llama 3.1 8B	0	33.1	52.2	0	68.6	56.8
Llama 3.1 70B	0	38.8	0.0	0	46.5	9.4
Mistral 7B	0	24.8	11.4	0	30.0	81.1
Mixtral 8x7B	0	9.2	29.2	0	79.2	39.8
SOLAR 10.7B	0	43.3	51.7	0	71.8	61.4
Yi 6B	0	12.5	19.8	0	72.7	71.6
Yi 34B	0	3.5	5.3	0	47.5	5.3
GPT-3.5 Turbo	0	63.6		0	63.6	
GPT-4o	0	16.0		0	18.0	

Table 21: Q&A with abstention results for WinoGrande. See Table 3 for an explanation of the scoring scheme.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0.1	0.1	0.1	−49.8	−8.7	−4.1
Falcon 40B	1.6	1.1	2.3	−47.6	−6.2	−6.3
Llama 2 7B	1.9	1.9	1.9	−47.1	−2.5	−0.5
Llama 2 70B	3.8	6.8	2.9	−44.3	−15.9	−20.8
Llama 3.0 8B	11.7	11.4	6.4	−32.5	−3.5	1.0
Llama 3.0 70B	49.7	46.9	47.0	24.6	34.0	25.9
Llama 3.1 8B	14.7	14.6	14.7	−27.9	−0.6	0.7
Llama 3.1 70B	55.7	54.2	51.9	33.5	40.4	27.3
Mistral 7B	1.7	5.8	6.0	−47.5	0.0	−1.0
Mixtral 8x7B	25.6	25.6	24.0	−11.6	−0.6	−1.8
SOLAR 10.7B	37.6	34.5	37.6	6.4	6.9	6.4
Yi 6B	17.4	17.4	17.4	−23.9	0.2	0.0
Yi 34B	35.0	35.0	35.0	2.5	2.5	2.5
GPT-3.5 Turbo	21.6	16.9		−17.7	5.9	
GPT-4o	63.8	63.8		45.8	45.8	

Table 22: Frequency of abstention on WinoGrande in the Section 6 experiments.

	Balanced			Conservative		
LLM	Base	MSP	Max Logit	Base	MSP	Max Logit
Falcon 7B	0	0.0	0.0	0	82.5	91.0
Falcon 40B	0	84.2	20.7	0	84.2	85.0
Llama 2 7B	0	0.0	0.0	0	93.7	98.5
Llama 2 70B	0	15.2	49.8	0	50.1	49.8
Llama 3.0 8B	0	21.5	76.0	0	67.7	93.5
Llama 3.0 70B	0	27.2	10.7	0	27.2	10.7
Llama 3.1 8B	0	1.5	0.0	0	68.8	99.0
Llama 3.1 70B	0	16.4	9.8	0	22.1	33.4
Mistral 7B	0	39.8	41.9	0	100.0	80.9
Mixtral 8x7B	0	0.0	13.4	0	78.3	49.6
SOLAR 10.7B	0	10.1	0.0	0	10.1	0.0
Yi 6B	0	0.0	0.0	0	98.8	100.0
Yi 34B	0	0.0	0.0	0	0.0	0.0
GPT-3.5 Turbo	0	61.1		0	61.1	
GPT-4o	0	0.0		0	0.0	