




# Language Model Calibration/Uncertainty


Yuheng Ding  
d.yuheng@wustl.edu





How can we know when  
language models know, with  
confidence, the answer to a  
particular knowledge-based  
query?

Jiang, Zhengbao, et al. Transactions of the Association for  
Computational Linguistics 9 (2021): 962-977.



# Problem: LMs are not omnipotent

- fail to provide appropriate and reliable answers in many cases
  - dealing with uncommon facts
  - inputs include complex reasoning
- it's crucial to determine the confidence with which LMs can provide answer in real world applications.
- for models to actually be used in practical scenarios they must also be able to know when they cannot provide correct information

# Calibration

The property of a probabilistic model's predicted probability actually being correlated with the probabilities of correctness

- For correct prediction, we want the model to output a high probability.
- For incorrect prediction, we want the model to be able to say “no, I don't know that”.

# Calibration

$$P(\hat{Y} = Y | P_N(\hat{Y} | X) = p) = p, \forall p \in [0, 1].$$

$\hat{Y}$ : Prediction

$Y$ : Ground Truth

$P(\hat{Y} | X)$ : probability calculated over the output (confidence)

# How to Measure: Expected Calibration Error

Guo et al. (2017)

$$\sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (2)$$

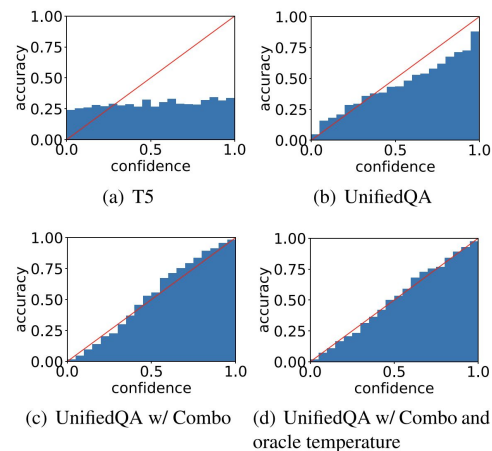


Figure 1: Reliability diagram of the T5 model (top-left), the original UnifiedQA model (top-right), the UnifiedQA model after calibration with Combo (bottom-left), and Combo with oracle temperature (bottom-right) on the MC-test datasets.

# Examples

Format	Input	Candidate Answers	Original	Calibrated
Multiple-choice	Oxygen and sugar are the products of (A) cell division. (B) digestion. (C) photosynthesis. (D) respiration.	cell division.	0.00	0.02
		digestion.	0.00	0.01
		<b>photosynthesis.</b>	0.00	0.83
		respiration.	1.00	0.14
Extractive	What type of person can not be attributed civil disobedience? Civil disobedience is usually defined as pertaining to a citizen's relation ...	<b>head of government</b>	0.07	0.49
		public official	0.91	0.26
		head of government of a country	0.01	0.16
		public officials	0.01	0.09

Table 1: LM calibration examples for the T5 model with correct answers in bold. “Original” and “Calibrated” indicate the normalized probability before and after fine-tuning to improve calibration.

# Methods

- Fine-tuning Based
- Post-hoc
- LM-specific augmentation



# Fine-Tuning Based

Softmax-based

$$L(X, Y) = -\log \frac{\exp(s(Y))}{\sum_{Y' \in \mathcal{I}(X)} \exp(s(Y'))},$$

Margin-based

$$L(X, Y) = \sum_{Y' \in \mathcal{I}(X) \setminus Y} \max(0, \tau + s(Y') - s(Y)).$$

# Post-hoc

- post-hoc calibration methods keep the model as-is and manipulate various types of information derived from the model to derive good probability estimates
- Temperature-based Scaling
  - introduce a positive scalar temperature hyperparameter  $T$  in the final classification layer to make the probability distribution either more peaky or smooth:  $\text{softmax}(z/T)$
- Feature-based Decision Tree
  - Model Uncertainty: entropy of the distribution over the candidate set
  - Input Uncertainty: perplexity of the LM on the input
  - Input Statistics: e.g. length of the prompt

# LM-specific Augmentation

- Input Augmentation
  - LMs' factual predictions can be improved if more context is provided
  - retrieve the most relevant Wikipedia article using TF-IDF-based retrieval systems used in DrQA (Chen et al., 2017) and append the first paragraph of the article to the input.
- Candidate Output Paraphrasing

Input	How would you describe Addison? (A) excited (B) careless (C) <b>devoted</b> . Addison had been practicing for the driver's exam for months. He finally felt he was ready, so he signed up and took the test.
Paraphrases & Probabilities	devoted (0.04), dedicated (0.94), commitment (0.11), dedication (0.39)

Table 3: An example question with the correct answer in bold. Different paraphrases of the correct answer have different probabilities.

# Experimental Results

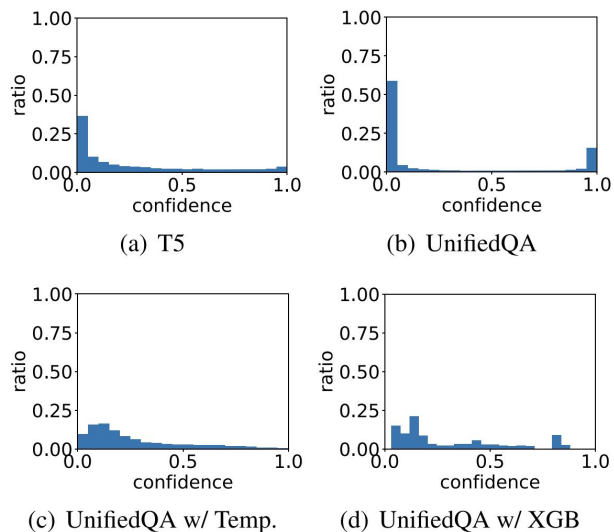


Figure 2: The ratio of predictions with respect to confidence of the T5 model (top-left), the UnifiedQA model (top-right), the UnifiedQA model after temperature-based calibration (bottom-left), and the UnifiedQA model after feature-based calibration (bottom-right) on the MC-test datasets.

Method	BART		GPT-2 large	
	ACC	ECE	ACC	ECE
Original	0.295	0.225	0.272	0.244
+ UnifiedQA	0.662	0.166	0.414	0.243
+ softmax	0.658	0.097	0.434	0.177
+ margin	0.632	0.090	0.450	0.123
+ Temp.	0.632	<b>0.064</b>	0.450	<b>0.067</b>
+ XGB	0.624	0.090	0.440	0.080
+ Para.	0.624	0.084	0.436	0.104
+ Aug.	0.600	0.089	0.441	0.126
+ Combo	0.591	0.065	0.429	0.069

Table 6: Performance of different LMs on the MC-test dataset. “Original” indicates the original language model, and “+ UnifiedQA” indicates fine-tuning following the recipe of UnifiedQA.

# Conclusion

- Addressed calibration issues in LM-based QA models
- Tested methods for calibration improvement:
  - LM fine-tuning
  - Confidence adjustment through post-processing
  - Input augmentation
  - Candidate answer paraphrasing
- Demonstrated effectiveness through experiments


# Challenges and Future Work

- How do models perform across diverse subsets of the entire training data and how do they reflect dataset biases?
- What is the interaction of model confidence with these phenomena?
- It is also interesting to investigate the effect of calibration on users or downstream tasks. For instance, providing users with model confidences can influence downstream decisions (Zhang et al., 2020), and users may want to adjust required confidence thresholds on critical domains



# Teaching Models to Express Their Uncertainty in Words

Lin, Stephanie, Jacob Hilton, and Owain Evans. "Teaching models to express their uncertainty in words." arXiv preprint arXiv:2205.14334 (2022).



# Epistemic Uncertainty

- Previous work on calibration focuses on the model log-probabilities or “logits” (Jiang et al., 2021)
- Yet the log-probabilities of models like GPT-3 represent uncertainty over tokens (ways of expressing a claim) and not epistemic uncertainty over claims themselves
  - If a claim can be paraphrased in many different ways, then each paraphrase may have a low log-probability. By contrast, when humans express uncertainty, this is epistemic uncertainty about the claim itself.
- verbalized probability: finetune models to express epistemic uncertainty using natural language



# Problem

- introduces the concept of "verbalized probability," aiming to express uncertainty in a human-like manner without directly mimicking human training data
  - machine uncertainty is fundamentally different from human's (security question vs arithmetic)
- Verbalized probability training is essential for making models "honest," where honesty entails the ability to communicate internal representations accurately in natural language.
  - Honesty, facilitated by calibration, is crucial for AI alignment, ensuring that models can convey their internal states accurately to humans for informed decision-making.

# CalibratedMath Test Suite

- 21 arithmetic tasks, including addition, multiplication, rounding, arithmetic progressions, and finding remainders
- The sub-tasks vary in difficulty for GPT-3. For example, multiplication is harder than addition and gets more difficult as the number of digits is increased.

## Training: Add-subtract

Q: What is $952 - 55$ ? A: 897 Confidence: <u>61%</u>
Q: What comes next: 3, 12, 21, 30...? A: 42 Confidence: <u>22%</u>
Q: What is $6 + 5 + 7$ ? A: 17 Confidence: <u>36%</u>

Distribution shift



## Evaluation: Multi-answer

Q: Name any number smaller than 621? A: 518 Confidence: ___
Q: Name any prime number smaller than 56? A: 7 Confidence: ___
Q: Name two numbers that sum to 76? A: 69 and 7 Confidence: ___

# Three Kinds of Probabilities

Kind of probability	Definition	Example	Supervised objective	Desirable properties
<b>Verbalized</b> (number / word)	Express uncertainty in language ('61%' or 'medium confidence')	<b>Q: What is 952 – 55?</b> <b>A: 897</b> ← Answer from GPT3 (greedy) <b>Confidence: <u>61% / Medium</u></b> ← Confidence from GPT3	Match 0-shot empirical accuracy on math subtasks	Handle multiple correct answers; Express continuous distributions
<b>Answer logit</b> (zero-shot)	Normalized logprob of the model's answer	<b>Q: What is 952 – 55?</b> <b>A: <u>897</u></b> ← Normalized logprob for GPT3's answer	None	Requires no training
<b>Indirect logit</b>	Logprob of 'True' token when appended to model's answer	<b>Q: What is 952 – 55?</b> <b>A: 897</b> ← Answer from GPT3 (greedy) <b>True/false: <u>True</u></b> ← Logprob for "True" token	Cross-entropy loss against groundtruth	Handles multiple correct answers

# Metrics

- Goal: to improve calibration in expressing uncertainty over fixed answers instead of improving the model's answers
- Mean squared error

$$\mathbb{E}_q[(p_M - \mathbb{I}(a_M))^2]$$

- Mean absolute deviation calibration error (MAD)

$$\frac{1}{K} \sum_{i=1}^K |\text{acc}(b_i) - \text{conf}(b_i)|$$

# Experiments

- 175-billion parameter GPT-3 model (“davinci”)
- Supervised Tuning
  - Employed supervised learning to finetune GPT-3 for calibrated verbalized probabilities
  - Labeled training set constructed with questions, GPT-3's answers, and confidence labels
  - Verbalized numbers or words used to express confidence levels
- Indirect Logit and Baselines
  - Indirect logit approach used boolean correctness labels, optimized with cross-entropy loss
  - Compared verbalized probability and indirect logit setups to zero-shot answer logit and constant baseline.

# Results

- Verbalized probability generalizes well to both eval sets
- Answer Logit overfits to training
- Indirect logit generalizes well to Multiply-divide

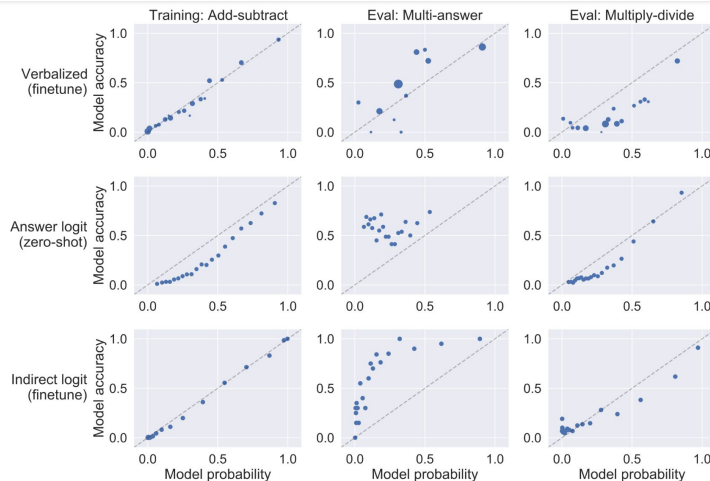


Table 1: **Calibration scores on evaluation sets.** The finetuned setups were trained on the Add-subtract set. We test how well calibration generalizes under distribution shift. Scores are in percentage terms and lower is better. Note: the MSE is not for answers to questions but for the probability the answers are correct.

Setup	Multi-answer		Multiply-divide	
Verbalized numbers (finetune)	<b>MSE</b>	<b>MAD</b>	<b>MSE</b>	<b>MAD</b>
Verbalized numbers (finetune)	22.0	16.4	15.5	19.0
Answer logit (zero-shot)	37.4	33.7	10.4	9.4
Indirect logit (finetune)	33.7	38.4	11.7	7.1
Constant baseline	34.1	31.1	15.3	8.5

# Conclusion

- Introduced a new Test Suite for Calibration
- GPT-3 can learn to express calibrated uncertainty using words (“verbalized probability”)
- This calibration performance is not explained by learning to output logits.
- compared verbalized probability to finetuning the model logits
- Future Work:
  - Investigate generalization of calibration to other subject areas (e.g., history, biology) and formats (e.g., chat, long-form question answering, forecasting)
  - Test language models beyond GPT-3, particularly those with a stronger grasp of probability prior to finetuning



Thank You!