Model Calibration

Emily Chang
Department of Computer Science
University of Virginia
Charlottesville, VA
ec5ug@virginia.edu

Jade Gregoire
Department of Computer Science
University of Virginia
Charlottesville, VA
dze3jz@virginia.edu

Michael Jerge
Department of Computer Science
University of Virginia
Charlottesville, VA
mj6ux@virginia.edu

What is a well-calibrated model?

A model's predicted probabilities for accuracy should be well-correlated with ground truth probabilities of correctness

Input	Candidate Answers	Original
Oxygen and sugar are the products of (A) cell division. (B) digestion. (C) photosynthesis. (D) respiration.	cell division.	0.00
	digestion.	0.00
	photosynthesis.	0.00
	respiration.	1.00

An example of a not-very-well calibrated model

How Can We Know *When* Language Models Know? On the Calibration of Language Models for Question Answering: https://arxiv.org/pdf/2012.00955.pdf

Common methods of assigning confidence

Probability (LM)

$$\operatorname*{argmax}_i P(\mathbf{y}_i|\mathbf{x})$$

Average Log-Likelihood (Avg)

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} P(y_i^j|\mathbf{x},\mathbf{y}^{1\cdots j-1})}{\ell_i}$$

Contextual Calibration (CC)

$$\arg\max_i \mathbf{w}P(\mathbf{y}_i|\mathbf{x}) + \mathbf{b}$$

Surface Form Competition: Why the Highest Probability Isn't Always Right https://arxiv.org/pdf/2104.08315.pdf

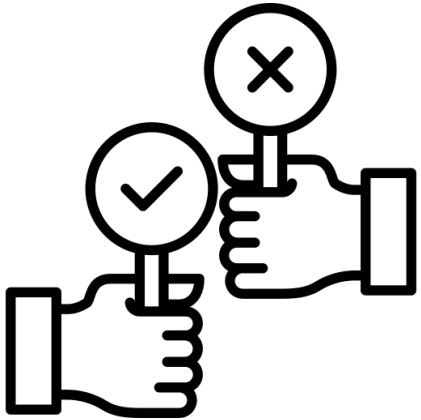
Form vs Meaning

“Different” sentences may be semantically equivalent to humans but models may be uncertain between two forms of the same meaning

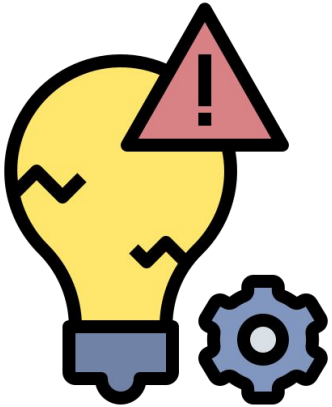
France’s capital is Paris.
Paris is the capital of France.

Why should you care?

Life Or Death



Decision Making



AI Alignment

Hallucinations

Image Source: https://www.flaticon.com/free-icon/mistake_4249097?term=mistake&page=1&position=4&origin=search&related_id=4249097

UNIVERSITY of VIRGINIA

How Can We Know *When* Language Models Know? On the Calibration of Language Models for Question Answering

Zhengbao Jiang[†], **Jun Araki**[‡], **Haibo Ding**[‡], **Graham Neubig**[†]
[†]Languages Technologies Institute, Carnegie Mellon University
[‡]Bosch Research
{zhengbaj,gneubig}@cs.cmu.edu
{jun.araki,haibo.ding}@us.bosch.com

UNIVERSITY of VIRGINIA

Overview

Research Question

- How can we know with confidence the answer to a particular query?

Results

- Determine that “strong” generative models are not well calibrated
- Methods to calibrate models are effective

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

B_m m-th bucket containing samples whose prediction confidence falls into interval $(\frac{m-1}{M}, \frac{m}{M}]$

$\text{acc}(B_m)$ average accuracy of m-th bucket

$\text{conf}(B_m)|$ average confidence of m-th bucket

Expected Calibration Error

weighted average of the discrepancy between each bucket's accuracy and confidence

UNIVERSITY of VIRGINIA

Calibration Methods

Fine-tuning: directly tune $P_N(\hat{Y}|X)$ to be a good probability estimate of actual answers Y

Post-hoc Calibration: manipulate information derived from the model

Language Model-Specific Methods

Fine-tuning

Softmax-based

Maximize the probability corresponding to the correct candidate

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

Y ground truth

$s(Y) = \log P_{\text{LM}}(Y|X)$ logit of the output Y

Margin-based

Maximize the confidence margin between ground truth and incorrect results

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

Post-hoc Calibration

Temperature-based Scaling

Temperature hyperparameter τ alters probability distribution of final classification layer

- τ → 0: largest logit receives most of the probability mass → less diverse outputs
- τ → ∞: uniform distribution → more diverse outputs

Feature-based Scaling

Model Uncertainty use entropy of the distribution over the candidate set $\mathcal{I}(X)$ to determine how uncertain the model is

Input Uncertainty high uncertainty indicates the input is “out-of-distribution”

Input statistics longer text may provide more information than shorter text

Notated as **XGB**

LM-Specific Methods

Candidate Output Paraphrasing

Round-trip translation model

1. Translate candidate output $Y' \in \mathcal{I}(X)$ into German
2. Generate a set of paraphrases by back-generating the German into English
3. Sum up the probability of all paraphrases to re-calculate the probability of all paraphrases

Input	How would you describe Addison? (A) excited (B) careless **(C) devoted**. 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)

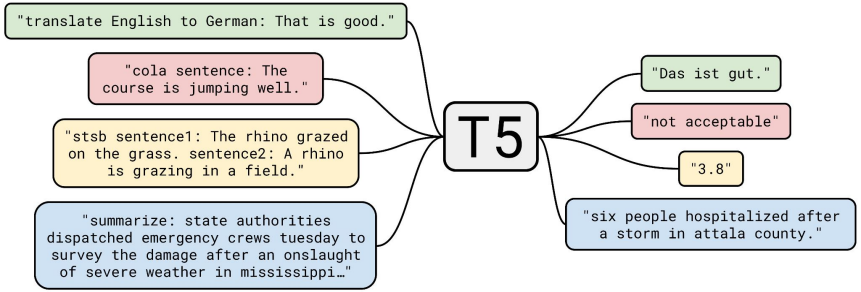
a candidate answer may not be worded in such a way that it achieves high confidence

LM-Specific Methods

Input Augmentation

- Retrieve extra evidence to augment input
- Find most relevant Wikipedia article and append first 3 sentences of the first paragraph

Models Evaluated



allenai/**unifiedqa**

UnifiedQA: Crossing Format Boundaries With a Single QA System

AI2

UNIVERSITY of VIRGINIA

Multiple-Choice QA

$$\hat{Y} = \arg\max_{Y' \in \mathcal{I}(X)} P_{\mathrm{LM}}(Y'|X)$$

X input

$\mathcal{I}(X)$ set of multiple choice answers

Y' a potential multiple choice answer

The answer that is returned is the highest-probability answer

Extractive QA

X

- Question
- Context passage containing answer to be extracted

Instead of calculating every possible span,

determine the top *K* spans as candidates and use candidates to calculate the probability (see MC QA)

$\mathcal{I}(X)$

UNIVERSITY of VIRGINIA

Datasets Used to Train and Evaluate

Format	Datasets and Domains
Multi-choice	ARC (science (Clark et al., 2018)), AI2 Science Questions (science (Clark et al., 2018)), OpenbookQA (science (Mihaylov et al., 2018)), Winogrande (commonsense (Sakaguchi et al., 2020)), CommonsenseQA (commonsense (Talmor et al., 2019b)), MCTest (fictional stories (Richardson et al., 2013)), PIQA (physical (Bisk et al., 2020)), SIQA (social (Sap et al., 2019)), RACE (English comprehension (Lai et al., 2017)), QASC (science (Khot et al., 2020)), MT-test (mixed (Hendrycks et al., 2020))
Extractive	SQuAD 1.1 (wikipedia (Rajpurkar et al., 2016)), SQuAD 2 (Wikipedia (Rajpurkar et al., 2018)), NewsQA (news (Trischler et al., 2017)), Quoref (wikipedia (Dasigi et al., 2019)), ROPES (situation understanding (Lin et al., 2019))

State-of-the-art models are not well-calibrated

Method	MC-test		MT-test		Ext-test	
	ACC	ECE	ACC	ECE	ACC	ECE
T5	0.313	0.231	0.268	0.248	0.191	0.166
UnifiedQA	0.769	0.095	0.437	0.222	0.401	0.114

Calibration can be achieved without sacrificing accuracy.

Method	MC-test ACC	MC-test ECE	MT-test ACC	MT-test ECE	Ext-test ACC	Ext-test ECE
T5	0.313	0.231	0.268	0.248	0.191	0.166
UnifiedQA	0.769	0.095	0.437	0.222	0.401	0.114
+ softmax	0.767	0.065	0.433	0.161	0.394	**0.110**
+ margin	0.769	**0.057**	0.431	**0.144**	0.391	0.112

Table 4: Performance of different fine-tuning methods.

Method	MC-test ACC	MC-test ECE	MT-test ACC	MT-test ECE	Ext-test ACC	Ext-test ECE
Baseline	0.769	0.057	0.431	0.144	0.401	0.114
+ Temp.	0.769	0.049	0.431	**0.075**	0.401	0.107
+ XGB	0.771	0.055	0.431	0.088	0.402	**0.103**
+ Para.	0.767	0.051	0.429	0.122	0.393	0.114
+ Aug.	0.744	0.051	0.432	0.130	0.408	0.110
+ Combo	0.748	**0.044**	0.431	0.079	0.398	0.104

Table 5: Performance of different post-hoc methods using the UnifiedQA model after margin-based fine-tuning or the original UnifiedQA model as the baseline model. "+Combo" denotes the method using both Temp., Para., and Aug.

Misplaced Confidence?

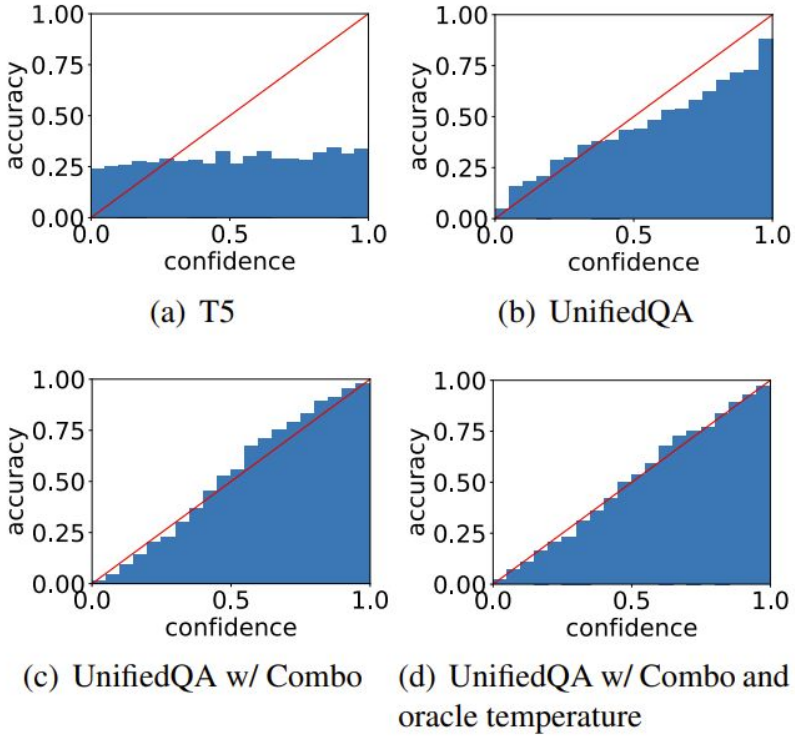


Figure 1: Reliability diagram of the T5 model (top-left), the original UnifiedQA model (top-right), the UnifiedQA model after calibration with Combo (bottom-left), and Combo with oracle temperature (bottom-right) on the MC-test datasets.

UNIVERSITY of VIRGINIA

Ablation Study

Calibrating Different Language Models

Method	MC-test ACC	MC-test ECE	MT-test ACC	MT-test ECE
T5	0.359	0.206	0.274	0.235
UnifiedQA	0.816	0.067	0.479	0.175
+ softmax	0.823	0.041	0.488	0.129
+ margin	0.819	0.034	0.485	0.107
+ Temp.	0.819	0.036	0.485	0.098
+ XGB	0.818	0.065	0.486	0.108
+ Para.	0.820	0.035	0.484	0.092
+ Aug.	0.812	**0.031**	0.493	0.090
+ Combo	0.807	0.032	0.494	**0.085**

Table 7: Performance of the 11B LMs.

Larger LMs achieve higher accuracy
better calibration results

Method	BART ACC	BART ECE	GPT-2 large ACC	GPT-2 large 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	**0.064**	0.450	**0.067**
+ 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.

Methods are applicable to LMs with
different architectures

Optimal number of paraphrases: 5-10

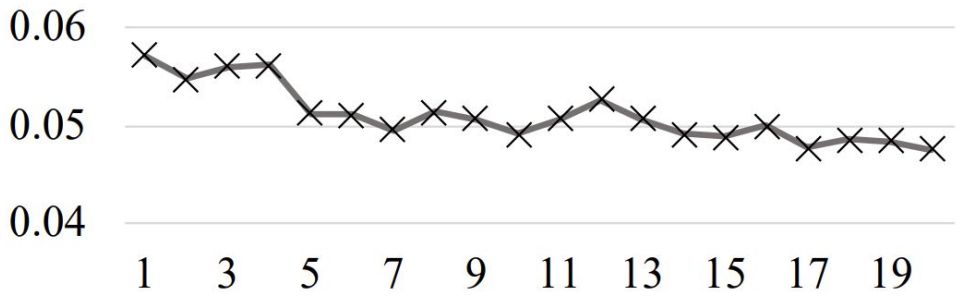


Figure 3: ECE of the UnifiedQA model using different numbers of paraphrases on the MC-test datasets.

ECE can generalize to out-of-domain datasets

Method	MC-train ACC	MC-train ECE	MC-test ACC	MC-test ECE
T5	0.334	0.228	0.313	0.231
UnifiedQA	0.727	0.133	0.769	0.095
+ softmax	0.735	0.084	0.767	0.065
+ margin	0.737	0.069	0.769	0.057
+ Temp.	0.737	0.051	0.769	0.049
+ XGB	0.737	0.074	0.771	0.055
+ Para.	0.742	0.053	0.767	0.051
+ Aug.	0.721	0.059	0.744	0.051
+ Combo	0.722	**0.042**	0.748	**0.044**

Table 8: Performance comparison between training and evaluation datasets.

Develop calibration methods on a more fine grained model

How does knowing the confidence affect users?

Surface Form Competition: Why the Highest Probability Answer Isn't Always Right

[=]Ari Holtzman[1] [=]Peter West[1,2]
Vered Shwartz[1,2] Yejin Choi[1,2] Luke Zettlemoyer[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
{ahai,pawest}@cs.washington.edu

UNIVERSITY of VIRGINIA

Overview

Issue

- Zero-shot capabilities of models are underestimated
- Ranking by string probability can be problematic, due to surface form competition

Solution

- Introduce Domain Conditional Pointwise Mutual Information

What is Surface Form Competition?

A human wants to submerge himself in water, what should he use?

Humans *select* options

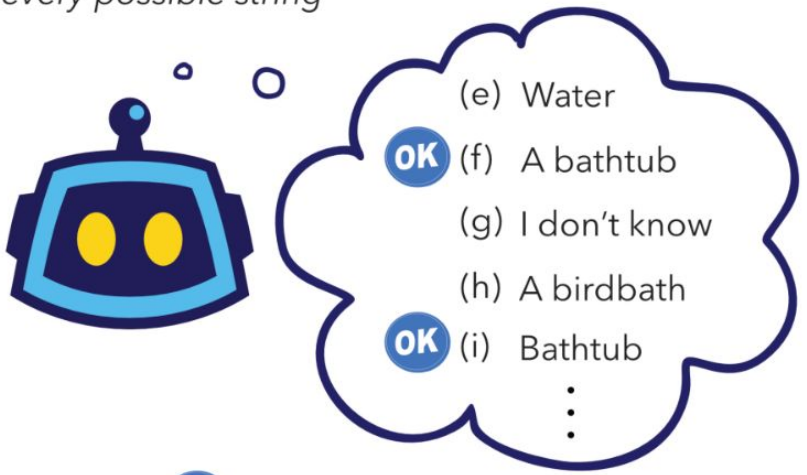
✗ (a) Coffee cup

✓ (b) Whirlpool bath

✗ (c) Cup

✗ (d) Puddle

Language Models assign probability to *every possible string*



OK = right concept, wrong surface form

Surface Form Competition: Why the Highest Probability Isn't Always Right
https://arxiv.org/pdf/2104.08315.pdf

UNIVERSITY of VIRGINIA

$$\underset{i}{\operatorname{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

Picking the highest-probability option: LM

Pointwise Mutual Information

$$\text{PMI}(\mathbf{x}, \mathbf{y}) = \log \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} = \log \frac{P(\mathbf{x}|\mathbf{y})}{P(\mathbf{x})}$$

- How much more likely does the hypothesis y become given the premise x
- Limitation: estimates of `P(y)` vary wildly

Domain Conditional Pointwise Mutual Information (PMI_{DC})

Reweighs scores by how much more likely a hypothesis (answer) becomes given a premise (question) within the specific task domain

Domain Conditional PMI (PMI_{DC})

$$\arg\max_i \frac{P(\mathbf{y}_i|\mathbf{x})}{P(\mathbf{y}_i|\mathbf{x}_{\text{domain}})}$$

Surface Form Competition: Why the Highest Probability Isn't Always Right https://arxiv.org/pdf/2104.08315.pdf

Existing Scoring Functions

Probability (LM)

$$\underset{i}{\operatorname{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

Average Log-Likelihood (AVG)

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} P(y_i^j|\mathbf{x},\mathbf{y}^{1\cdots j-1})}{\ell_i}$$

Unconditional (in-domain) estimate (UNC)

$$\arg\max_i P(\mathbf{y}_i|\mathbf{x}_{\text{domain}})$$

Setup

Models

- GPT-2 (via the HuggingFace Transformers library)
- GPT-3 (via OpenAI's beta API)

Datasets

- 13 datasets

Experiments

- Multiple Choice
  - Zero-shot (main focus)
  - Few-shot
- Removing Surface Form Competition
  - Scoring-by-premise

Datasets

- **Continuation**: Choice of Plausible Alternatives (COPA), StoryCloze (SC), HellaSwag (HS)
- **Question Answering (QA)**: RACE-M & -H (R-M & R-H), ARC Easy & Challenge (ARC-E & ARC-C), Open Book Question Answering (OBQA), CommonsenseQA (CQA)
- **Boolean QA**: BoolQ (BQ)
- **Entailment**: Recognizing Textual Entailment (RTE), Commitment Bank (CB)
- **Text Classification**: SST-2 & -5, AG's News, TREC

Results of Multiple Choice Experiments - Zero-shot

- Smallest margin: >40%, between AVG and PMI_{DC}
- Greatest margin: >80%
- PMI_{DC} performs significantly better on new datasets

Percent of Ties or Wins by Method

Method	Unc	LM	Avg	PMI_{DC}	CC
125M	12.50	6.25	12.50	**68.75**	-
350M	6.25	18.75	12.50	**68.75**	-
760M	6.25	6.25	12.50	**75.00**	-
1.6B	6.25	12.50	12.50	**80.00**	20.00
2.7B	6.25	6.25	6.25	**86.66**	0.00
6.7B	6.25	25.00	25.00	**75.00**	-
13B	6.25	18.75	18.75	**68.75**	-
175B	6.25	12.50	18.75	**62.50**	6.25

Surface Form Competition: Why the Highest Probability Isn't Always Right https://arxiv.org/pdf/2104.08315.pdf

UNIVERSITY of VIRGINIA

Results of Multiple Choice Experiments - Few-shot

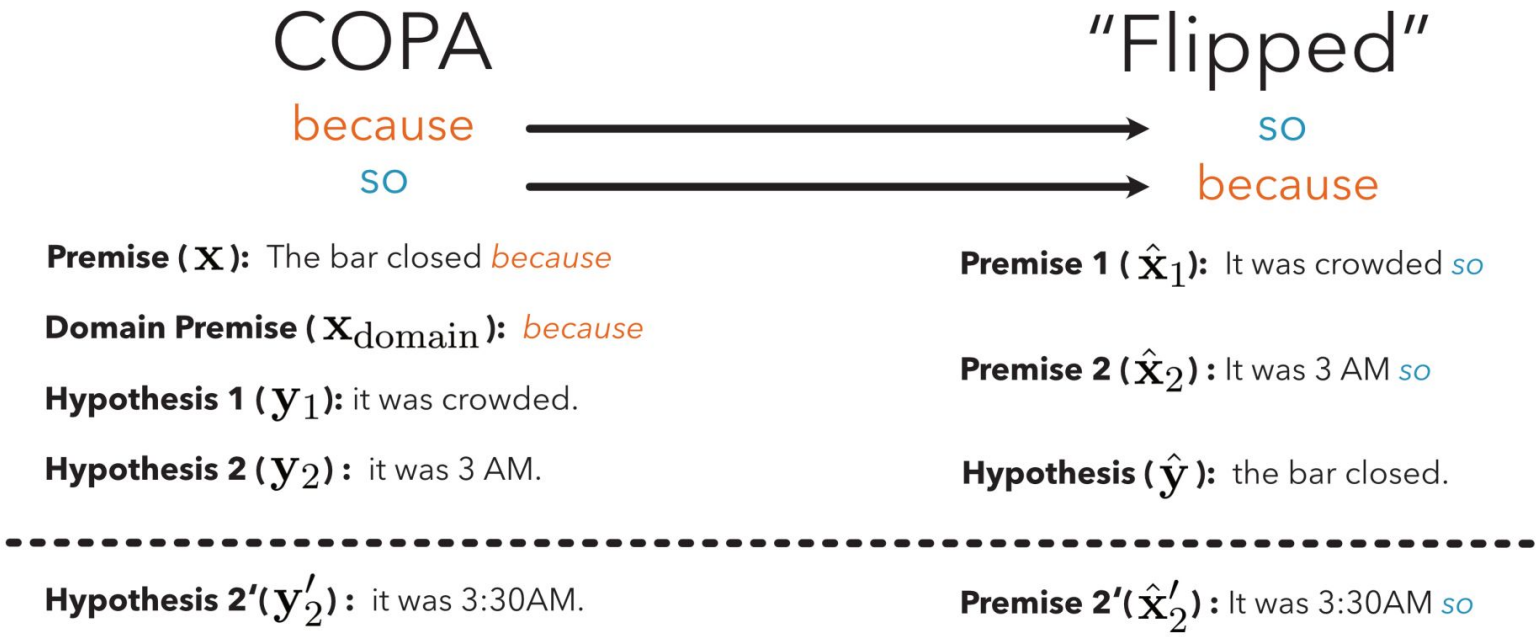
- PMI_{DC} favored
- LM performs better for two models on SST-2 dataset

4-shot Inference Results

	SST-2			CQA			
Method	Unc	LM	PMI_{DC}	Unc	LM	Avg	PMI_{DC}
125M	49.9 $_{0}$	63.6 $_{7.4}$	**71.7** $_{5.1}$	15.5 $_{0}$	29.9 $_{1.6}$	32.7 $_{1.4}$	**38.3** $_{1.7}$
350M	49.9 $_{0}$	76.3 $_{13.8}$	**76.4** $_{8.1}$	16.5 $_{0}$	37.6 $_{2.3}$	40.4 $_{2.3}$	**45.7** $_{2.4}$
760M	49.9 $_{0}$	85.9 $_{7.2}$	**87.1** $_{3.0}$	16.1 $_{0}$	41.5 $_{2.6}$	42.4 $_{2.5}$	**47.0** $_{1.5}$
1.6B	49.9 $_{0}$	85.4 $_{1.7}$	**89.4** $_{4.0}$	16.0 $_{0}$	46.2 $_{1.5}$	47.7 $_{1.9}$	**52.3** $_{2.1}$
2.7B	49.9 $_{0}$	**88.1** $_{4.9}$	87.7 $_{5.5}$	16.6 $_{0}$	43.0 $_{1.7}$	45.6 $_{1.9}$	**50.4** $_{1.1}$
6.7B	49.9 $_{0}$	**92.9** $_{2.1}$	79.8 $_{6.9}$	16.9 $_{0}$	52.3 $_{1.4}$	53.4 $_{1.0}$	**56.5** $_{1.6}$
13B	49.9 $_{0}$	85.4 $_{9.0}$	**86.9** $_{7.5}$	16.7 $_{0}$	58.4 $_{2.0}$	59.3 $_{1.5}$	**63.4** $_{1.4}$
175B	49.9 $_{0}$	89.9 $_{5.5}$	**95.5** $_{0.7}$	16.5 $_{0}$	69.1 $_{1.9}$	69.4 $_{0.8}$	**72.0** $_{0.9}$

Surface Form Competition: Why the Highest Probability Isn't Always Right
https://arxiv.org/pdf/2104.08315.pdf

Removing Surface Form Competition Experiment



Surface Form Competition: Why the Highest Probability Isn't Always Right https://arxiv.org/pdf/2104.08315.pdf

Scoring-by-Premise

$$P(\mathbf{x}|\mathbf{y})$$

Probability of the premise given the hypothesis

Eliminates competition from surface form by calculating the same surface form across different options

P(It was 3 AM | The bar closed)

P(It was crowded | The bar closed)

Only one answer is selected

P(The bar closed | It was 3 AM)

P(The bar closed | It was crowded)

Multiple answers could be selected

UNIVERSITY of VIRGINIA

Removing Surface Form Competition Results

- UNC produces the exact same results
- On COPA Flipped, LM/AVG perform similarly to PMI_{DC} on the unflipped version

Removing Surface Form Competition

	COPA				COPA Flipped			
Method	Unc	LM	Avg	PMI_{DC}	Unc	LM	Avg	PMI_{DC}
125M	56.4	61.0	63.2	62.8	50.0	63.2	63.2	63.2
350M	55.8	67.0	66.0	70.0	50.0	66.4	66.4	66.4
760M	55.6	69.8	67.6	69.4	50.0	70.8	70.8	70.8
1.6B	56.0	69.0	68.4	71.6	50.0	73.0	73.0	73.0
2.7B	54.8	68.4	68.4	74.4	50.0	68.4	68.4	68.4
6.7B	56.4	75.8	73.6	77.0	50.0	76.8	76.8	76.8
13B	56.6	79.2	77.8	84.2	50.0	79.0	79.0	79.0
175B	56.0	85.2	82.8	89.2	50.0	83.6	83.6	83.6

Surface Form Competition: Why the Highest Probability Isn't Always Right
https://arxiv.org/pdf/2104.08315.pdf

Why Does Scoring-By-Premise Work?

COPA $P(\mathbf{y}_1|\mathbf{x}) > P(\mathbf{y}_2|\mathbf{x})$

P(It was crowded | The bar closed) > P(It was 3 AM | The bar closed) ✗

COPA Flipped

$$P(\hat{\mathbf{y}}|\hat{\mathbf{x}}_2) > P(\hat{\mathbf{y}}|\hat{\mathbf{x}}_1)$$

$$\frac{P(\mathbf{y}_2|\mathbf{x})}{P(\mathbf{y}_2|\mathbf{x}_{\text{domain}})} > \frac{P(\mathbf{y}_1|\mathbf{x})}{P(\mathbf{y}_1|\mathbf{x}_{\text{domain}})}$$

P(The bar closed | It was 3 AM) > P(The bar closed | It was crowded) ✓

Stability over Multiple Answers

$$\log P(\mathbf{y}_2|\mathbf{x}) \approx -16$$ **P(it was 3 AM | the bar closed because)**

$$\log P(\mathbf{y}'_2|\mathbf{x}) \approx -20$$ **P(it was 3:30 AM | the bar closed because)**

Because the conditional probability for y_2' is lower than y_2, the score for **y_2 = it was 3 AM** will be different from **y_2' = it was 3:30 AM**

Stability over Multiple Answers: Scoring-by-Premise

$$\log P(\hat{\mathbf{y}}|\hat{\mathbf{x}}_2) \approx -12$$
$$\log P(\hat{\mathbf{y}}|\hat{\mathbf{x}}'_2) \approx -12$$

P(The bar closed | It was 3 AM so) P(The bar closed | It was 3:30 AM so)

Stabilizes the conditional probability of $\hat{\mathbf{y}}$

Conclusions, Limitations, & Future Work

- PMI_{DC} outperforms previous scoring functions on multiple choice datasets
  - Prove that this is due to surface form competition by showing how other scoring methods have improved accuracy when surface form competition is removed
- Limited by ability to understand answer concepts
  - In multiple choice, would not understand multiple answers that interact with each either, such as “all of the above”
- Should explore how surface form competition affects answer generation—may cause generic outputs when models are highly uncertain

Teaching models to express their uncertainty in words

Stephanie Lin *sylin07@gmail.com*
University of Oxford

Jacob Hilton *jhilton@openai.com*
OpenAI

Owain Evans *owaine@gmail.com*
University of Oxford

UNIVERSITY of VIRGINIA

Overview

Significance

- GPT-3 model can learn to express uncertainty about its own answers in natural language

Results

- Remains moderately well-calibrated under a distribution shift
- Is sensitive to uncertainty in its own answers, rather than using human examples

Potential Explanation

- GPT-3 uses latent (pre-existing) representations to generalize calibration

Why do models need to be truthful?

- Curbing hallucinations
- Previous research on using logits to represent uncertainty
  - Limited by calculation of uncertainty over tokens, not semantic meaning
- Self-awareness of misinformation or doubt in a model leads to better communication with users

Zero-Shot Setup

- **Model**: 175B parameter GPT-3 via OpenAI API
- **Metrics**:
  - Mean squared error (MSE) $\mathbb{E}_q[(p_M - \mathbb{I}(a_M))^2]$
  - Mean absolute deviation calibration error (MAD) $\frac{1}{K}\sum_{i=1}^{K}|\mathrm{acc}(b_i) - \mathrm{conf}(b_i)|$
- **Experiment**: Test calibration of language models for uncertainty over their own answers to questions with 3 different kinds of probability

Dataset: CalibratedMath

Training: Add-Subtract

Q: What is 952 - 55?
A: 897
Confidence: 61%

Evaluation: Multi-Answer

Q: Name any number smaller than 621
A: 518
Confidence: ___

Evaluation: Multiply-Divide

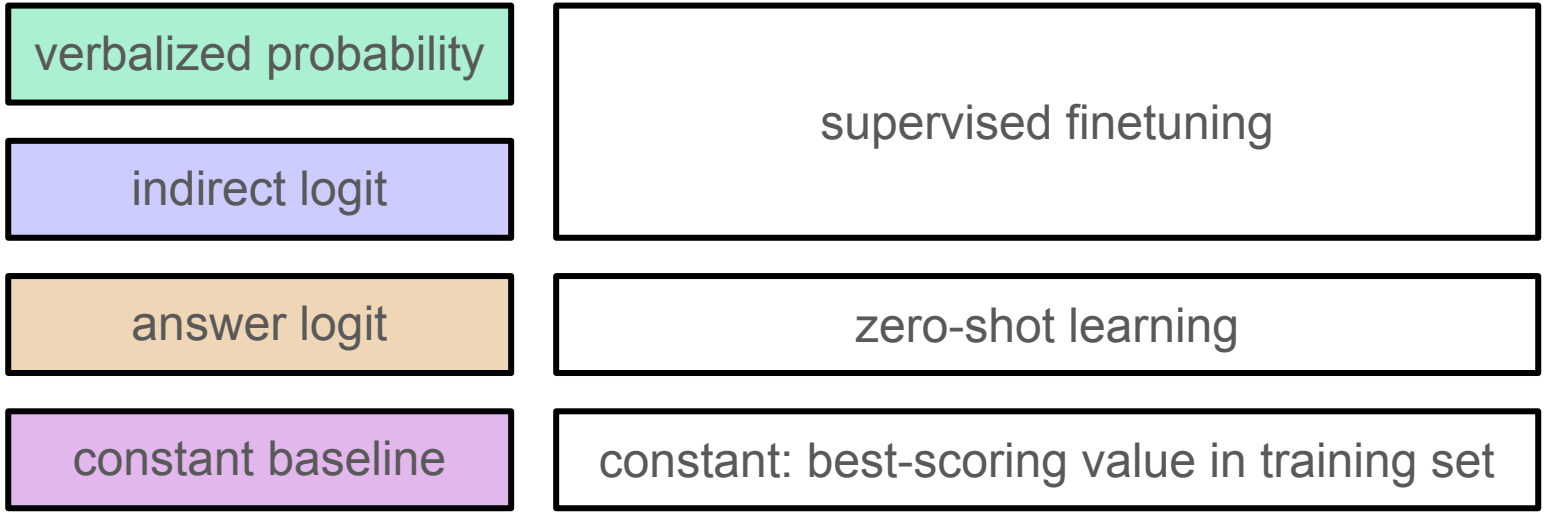
Q: What is 1111 × 1111
A: 123456789
Confidence: ___

Three Kinds of Probability

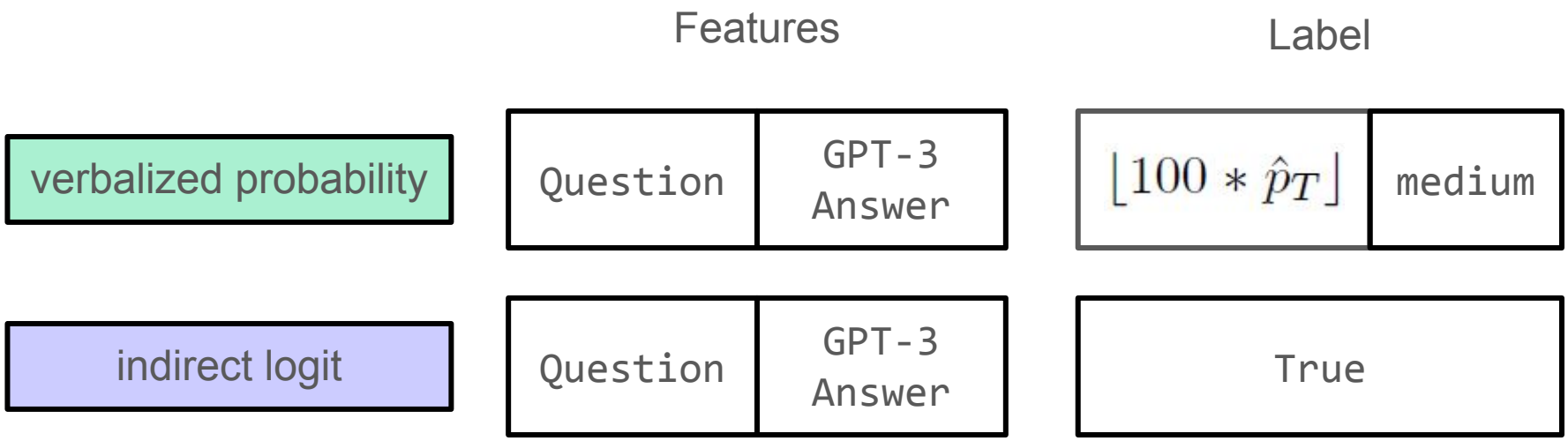
Kind of probability	Definition	Example	Supervised objective	Desirable properties
Verbalized (number / word)	Express uncertainty in language ('61%' or 'medium confidence')	**Q: What is 952 – 55?** **A: 897** ← Answer from GPT3 (greedy) **Confidence:** 61% / Medium ← Confidence from GPT3	Match 0-shot empirical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

Teaching models to express their uncertainty in words: https://arxiv.org/pdf/2205.14334.pdf

Implementing Verbalized Probability and Baselines



Supervised Finetuning

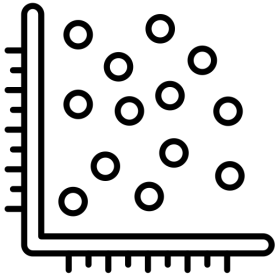


But, what if GPT-3 isn't good at math?

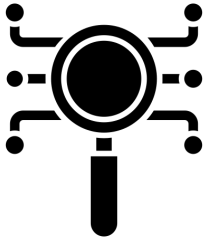
$10 \times 10 = 100$

GPT-3 finds two-digit multiplications hard → internal bias

For a task T



sample 100 data points



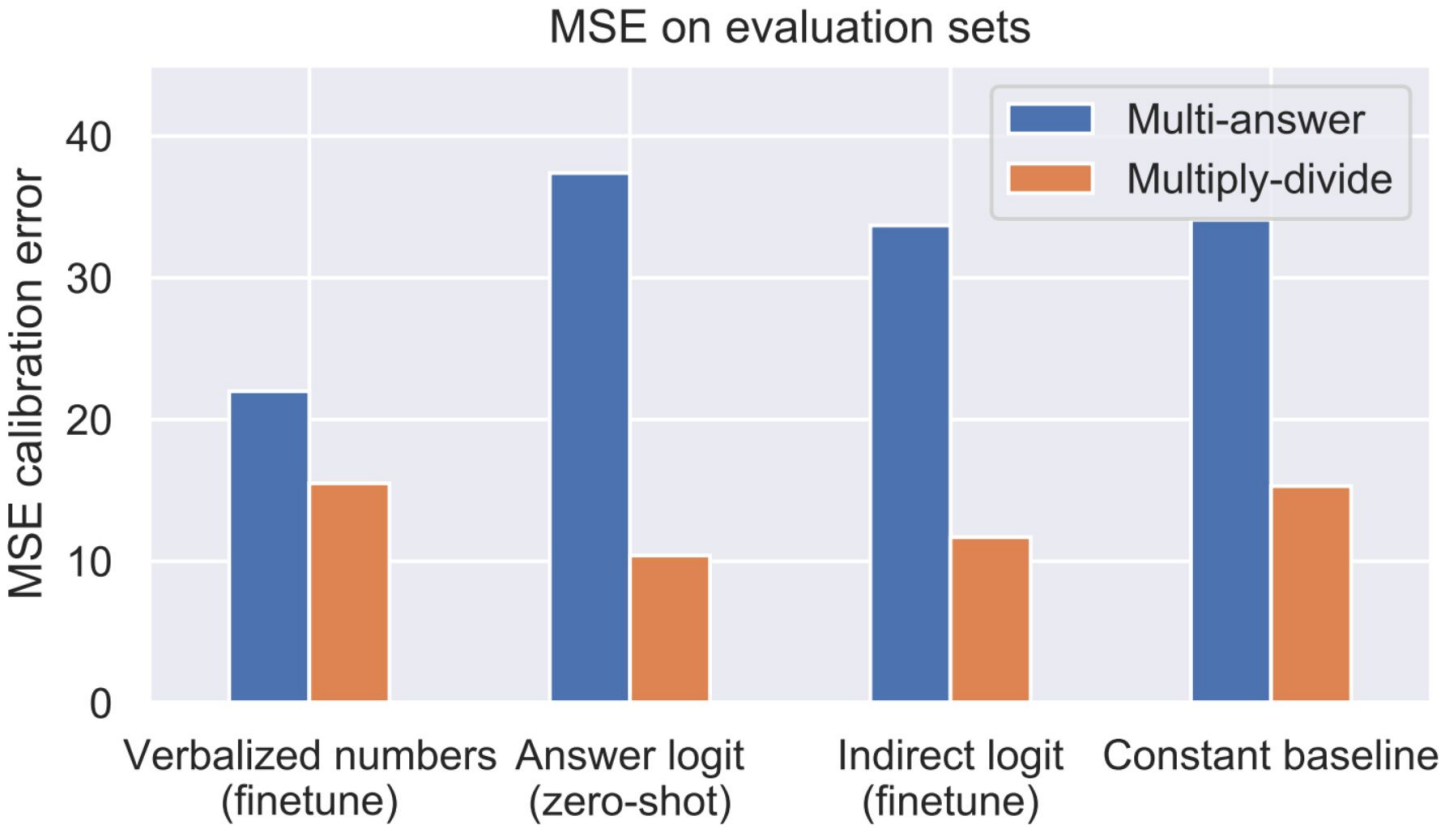
Retrieve answer using greedy-based decoding

$$\hat{p}_T = \mathbb{E}_{q \in T}[\mathbb{I}(a_M)]$$

align confidence score $\hat{p}_T$ to accuracy of the answer

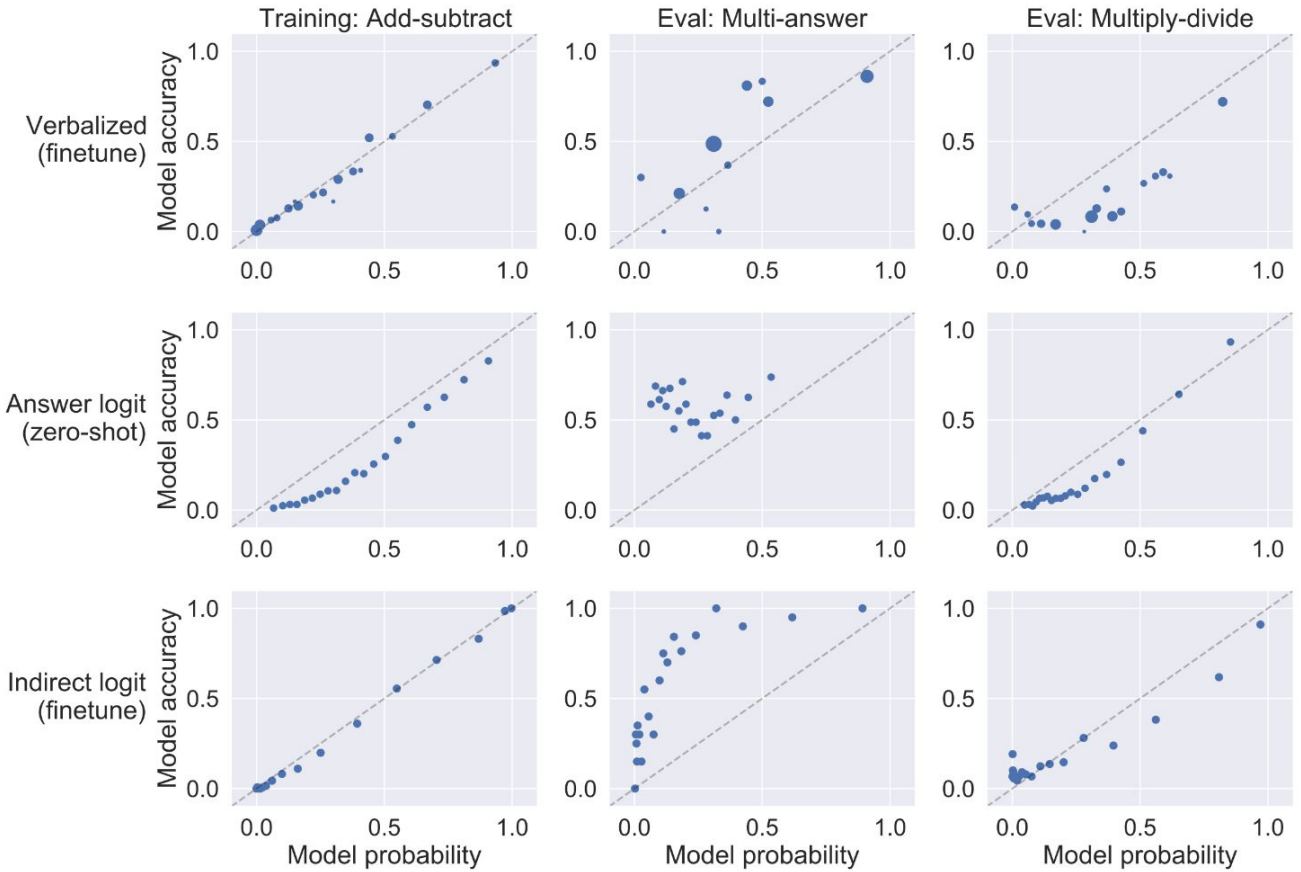
use $\hat{p}_T$ as label

Results



Teaching models to express their uncertainty in words: https://arxiv.org/pdf/2205.14334.pdf

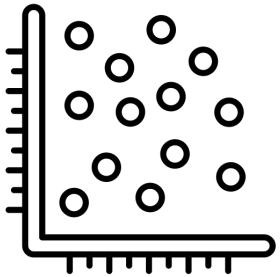
Results



Teaching models to express their uncertainty in words: https://arxiv.org/pdf/2205.14334.pdf

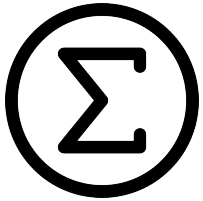
Stochastic Few Shot

Purpose: how does verbalized probability generalize?

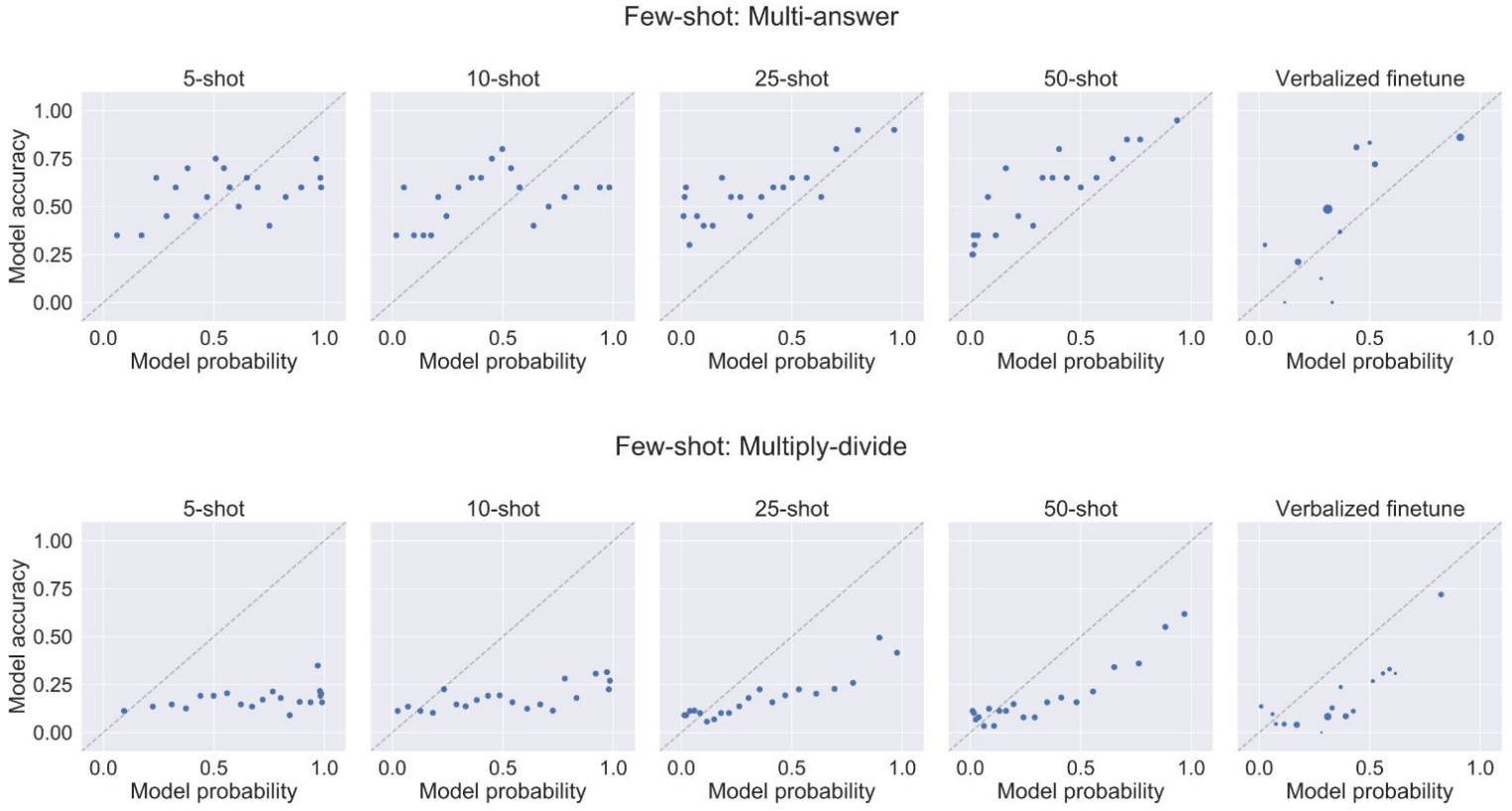


Sample k examples from the training set

Add to context



Expected value decoding



Very uncalibrated

Signs of calibration

Teaching models to express their uncertainty in words: https://arxiv.org/pdf/2205.14334.pdf

How does verbalized probability work?

- Verbalized probability generalizes better, does not rely on logits
- Verbalized probability cannot be fully explained by heuristics
- Model expresses its own (pre-existing) uncertainty about answers and exhibits honesty
    - Latent representations

Limitations & Future Work

- Jiang et al (see 1st paper)'s calibration is more expensive
- Paper focused only on a mathematical dataset—benefit from exploring other subject areas
- Expand to see if results are similar with other question formats
- Test with other models—not just GPT-3
- Explore other forms of learning, such as reinforcement learning (so that fine-tuning does not have to be supervised, and can use less resources)

SEMANTIC UNCERTAINTY: LINGUISTIC INVARIANCES FOR UNCERTAINTY ESTIMATION IN NATURAL LANGUAGE GENERATION

Lorenz Kuhn, Yarin Gal, Sebastian Farquhar
OATML Group, Department of Computer Science, University of Oxford
lorenz.kuhn@cs.ox.ac.uk

UNIVERSITY of VIRGINIA

Overview

Issue

- Because of semantic equivalence, measuring uncertainty in natural language is challenging

Solution

- Developed an unsupervised single model method that calculates semantic entropy
- Semantic entropy is more predictive of model accuracy for question answering

Formalizing Semantic Equivalence

$$\forall s, s' \in c : E(s, s')$$

For the space of semantic equivalence classes C the sequences in the set $c \in C$ all share a meaning under the semantic equivalence relation $E(\cdot, \cdot)$

Paris ⇏ Paris is the capital of France

Datasets

CoQA

Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up in a nice warm place above the barn where all of the farmer's horses slept. But Cotton wasn't alone in her little home above the barn, oh no. She shared her hay bed with her mommy and 5 other sisters. All of her sisters were cute and fluffy, like Cotton. But she was the only white one in the bunch. The rest of her sisters were all orange with beautiful white tiger stripes like Cotton's mommy. Being different made Cotton quite sad. She often wished she looked like the rest of her family. So one day, when Cotton found a can of the old farmer's orange paint, she used it to paint herself like them. When her mommy and sisters found her they started laughing.

Q: What color was Cotton?

A: white || a little white kitten named Cotton

TriviaQA

What was the Elephant Man's real name?	Joseph Merrick

Unsupervised Algorithm

Generating a set of answers from the model

- Sample *M* sequences $\{s^{(1)}, \ldots, s^{(m)}\}$ according to the distribution $p(s|x)$
- Performed using a single model

Up Next

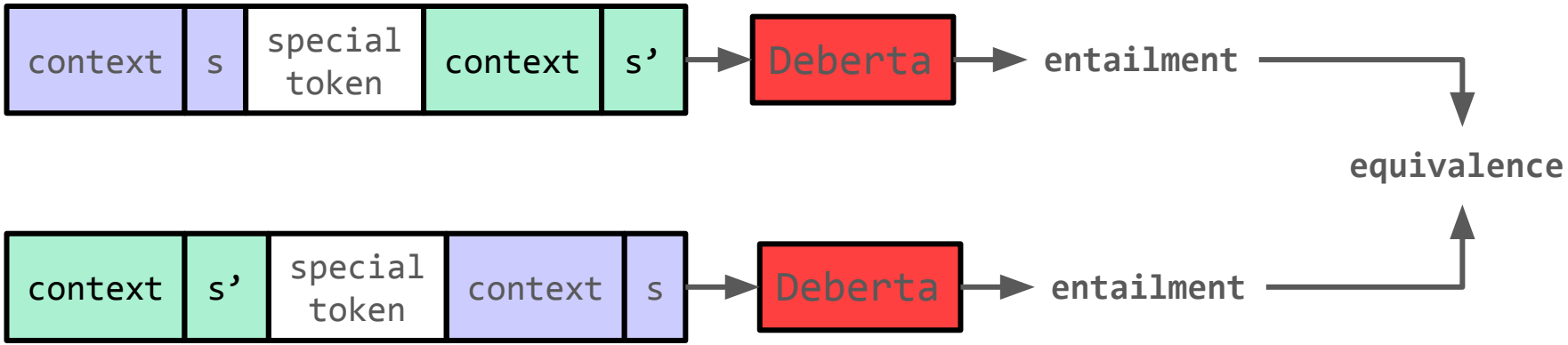
Clustering

Computing entropy

UNIVERSITY of VIRGINIA

Clustering by semantic equivalence

A sequence s means the same thing as a second sequence s' if and only if they entail each other.



CoQA

95.5% accuracy

TriviaQA

92.7% accuracy

Computing the semantic entropy

clusters of generated sequences that mean the same thing

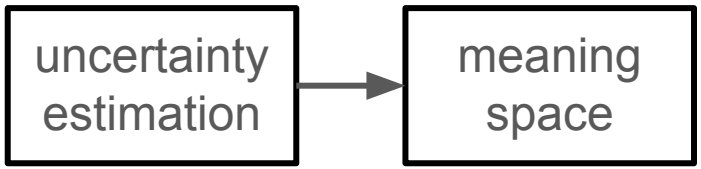
$$p(c \mid x) = \sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x) = \sum_{\mathbf{s} \in c} \prod_{i} p(s_i \mid s_{<i}, x)$$

Determine the likelihood of each meaning rather than each sequence

$$SE(x) \approx -|C|^{-1} \sum_{i=1}^{|C|} \log p(C_i \mid x)$$

Compute the semantic entropy

Semantic Entropy addresses…



semantic invariance of natural language

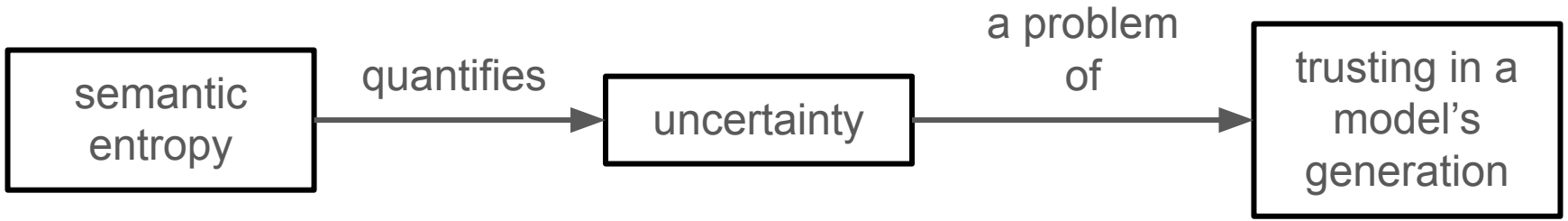
France's capital is Paris.
Paris is the capital of France

unequal token importance

Shortcomings of Semantic Entropy

Semantic entropy pays too much attention to non-keyword likelihoods

Potentially resolved by supervised language models


semantic entropy
quantifies
uncertainty
a problem of
trusting in a model's generation

Other Entropy Metrics

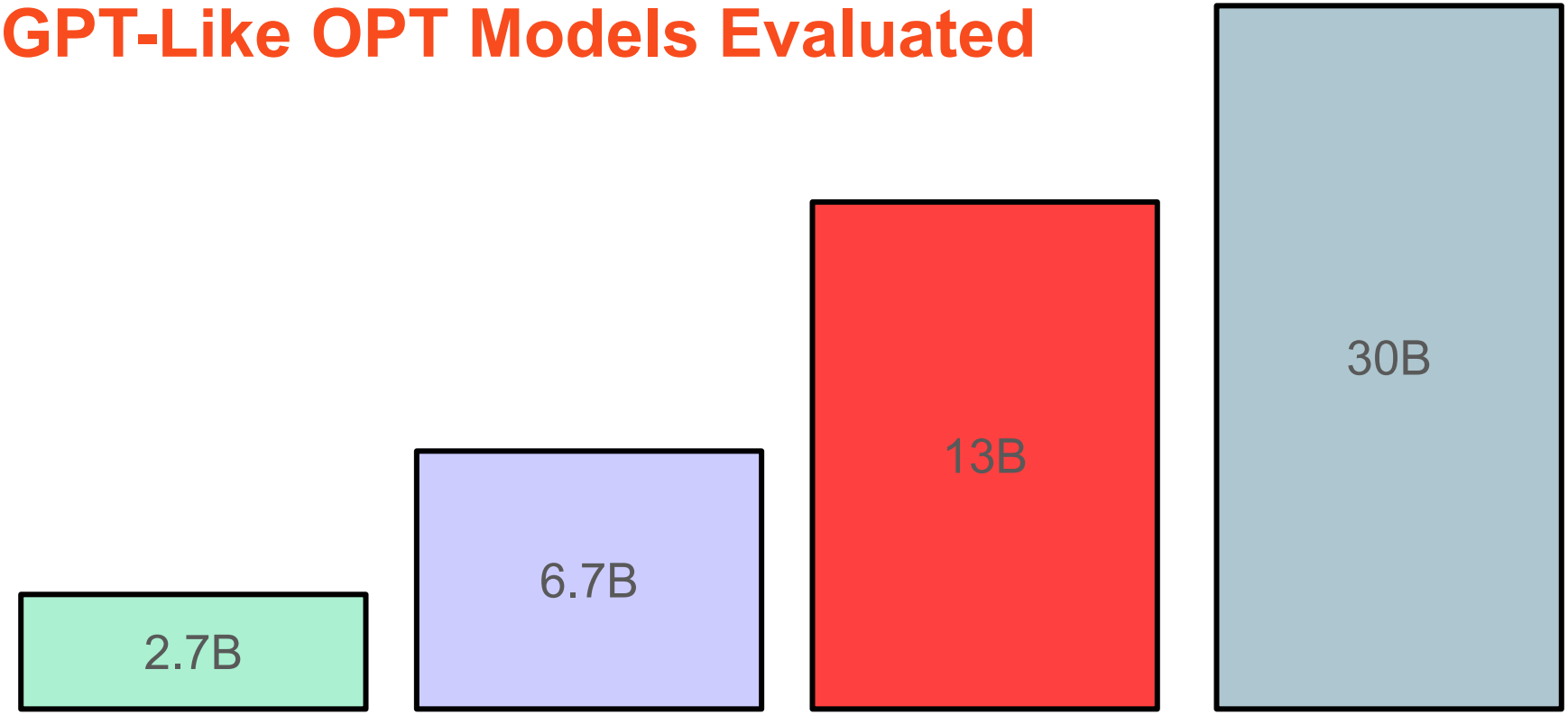
Normalised entropy: divides the joint log-probability of each sequence by the length of the sequence

(Predictive) entropy: conditional entropy of the output random variable Y with realization y given x $PE(x) = H(Y \mid x) = -\int p(y \mid x) \ln p(y \mid x) dy$

Lexical similarity: average similarity of the answer set $\mathbb{A}$: $\frac{1}{C} \sum_{i=1}^{|\mathbb{A}|} \sum_{j=1}^{|\mathbb{A}|} \text{sim}(s_i, s_j)$

p(True): “ask” the model if its answer is correct

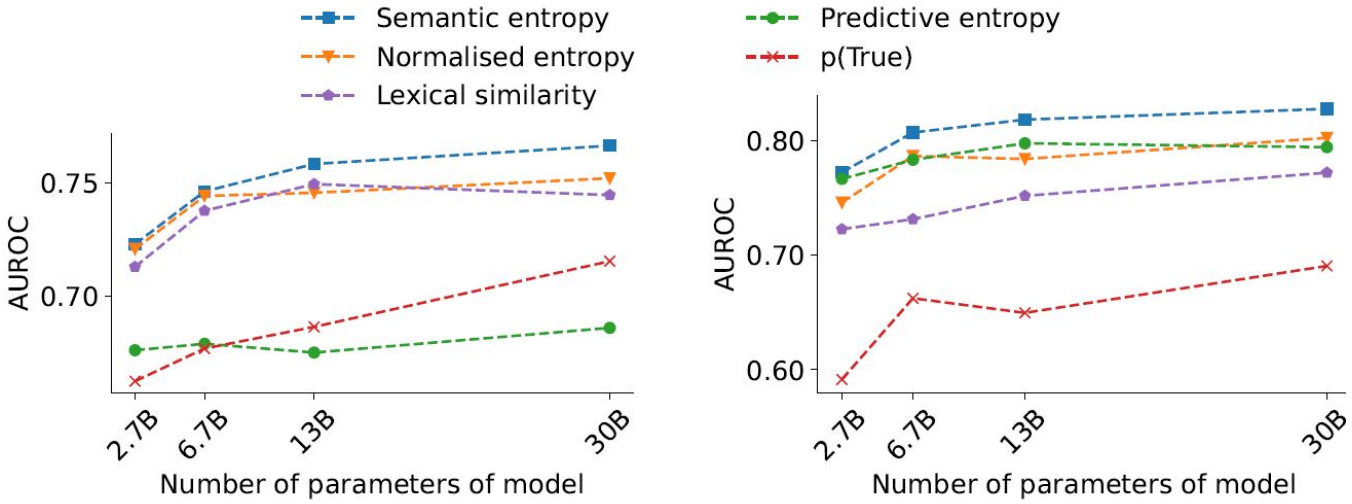
GPT-Like OPT Models Evaluated



UNIVERSITY of VIRGINIA

Area Under the Receiver Operator (AUROC)

- Equivalent to the probability a randomly chosen correct answer has a higher uncertainty score than a randomly chosen incorrect answer
- Higher the score, the better
- AUROC doesn't require probability mass → good metric for natural language generation

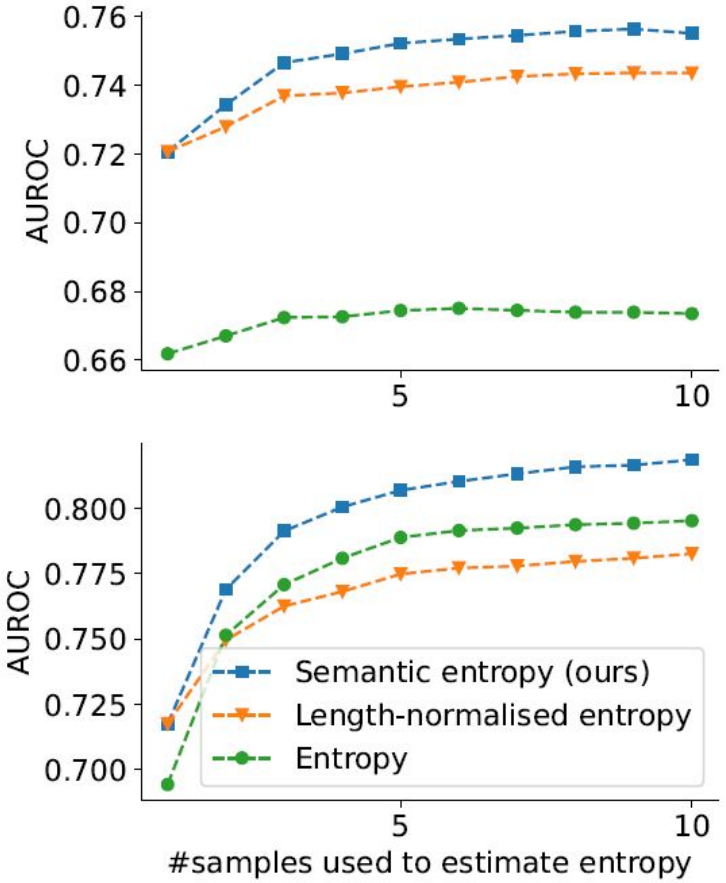


(a) CoQA (b) TriviaQA

Semantic entropy improves over baselines in predicting whether a model's answer to a question is correct.

Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation: https://arxiv.org/pdf/2302.09664.pdf

Semantic entropy makes better use of additional samples because it handles duplication better



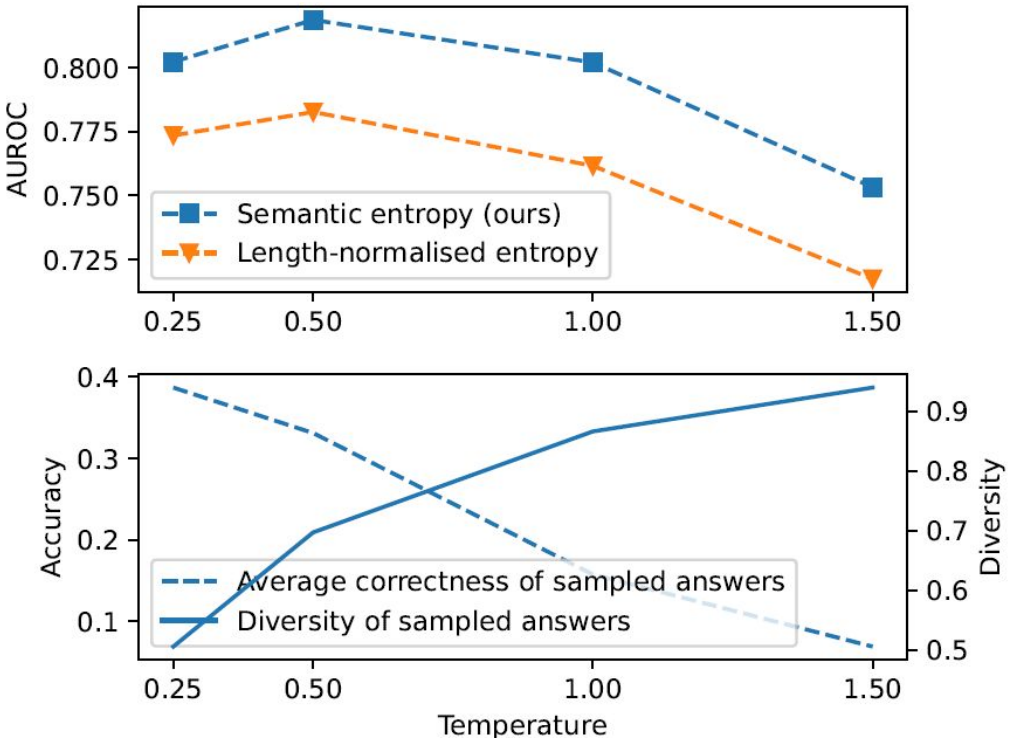
(a) (top) CoQA, (bottom) TriviaQA

Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation: https://arxiv.org/pdf/2302.09664.pdf

Temperature

used to control randomness and creativity of a language model

The best uncertainty comes from balancing diversity and accuracy



UNIVERSITY of VIRGINIA

Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation: https://arxiv.org/pdf/2302.09664.pdf

Limitations & Future Work

- Language models are capable of deception
  - The paper's method does not protect against this
  - Has potential to be added on to, to mitigate deception
- Pave the way towards progress in other NLG settings
  - Summarization
  - Reasoning

So, what did we learn about model calibration?

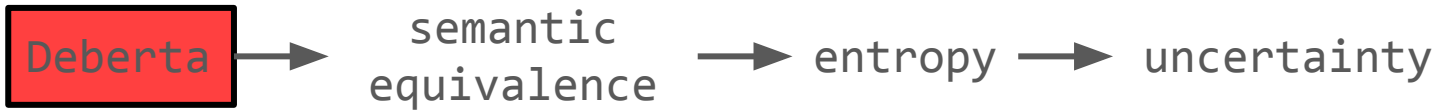
Confronting Semantic Equivalence

Jiang et al. (see 1st paper) and Schwartz et al.(see 2nd paper) attempt to take into account semantic equivalence

- Jiang et al's Combo method: round-trip translational model produces synonyms
- Schwartz et al: use the input x to re-rank the outputs y

Lin et al. (see 3rd paper) does not even account for semantic equivalence

Kuhn et al. (see 4th paper) link semantic equivalence to calibration



Well-Calibrated models

If the answer is correct, the model should be highly confident in its answer

Challenge: Form vs. Meaning

Models can fail to recognize when sequences of tokens mean the same thing, affecting calibration

Detecting meaning instead of form

Entropy and unsupervised learning methods can detect semantically equivalent sequences. Entropy can be a metric for the model's confidence

Moving towards better calibration

Calibration methods such as input-augmentation and PMI_{DC} have proven effective. Calibration as natural language can make this field of study more accessible to non-technical users

Thank you!

Any Questions?