Uncertainty and Calibration

Lecture 14

CS 580B1 / CS 480B1, Trustworthy AI, Fall 2026

Ravi Mangal

Roadmap

	Section	
1	Confidence and calibration	classifiers
2	Abstention	classifiers
3	Distribution shift	classifiers
4	Language models: multiple choice	one token
5	Language models: free-form answers	sequences and meanings

1

Confidence and calibration

Confidence

$$\hat{y}(x) = \arg\max_k p_k(x) \qquad \hat{p}(x) = \max_k p_k(x)$$

prediction and confidence, from p(x) = softmax(z(x))

Uncertainty

$$H(p(x)) = -\sum_k p_k(x) \log p_k(x)$$

- **Uncertainty**: how spread out p(x) is. Measures: 1 − p̂(x), entropy H.
- **Aleatoric** (from the input) vs **epistemic** (from what the model has not learned)
- **Uncertainty measure**: any score meant to be high when the prediction is likely wrong

Calibration

$$P_{(x,\,y)\,\sim\,D}(\hat{y}(x) = y \mid \hat{p}(x) = c) = c$$

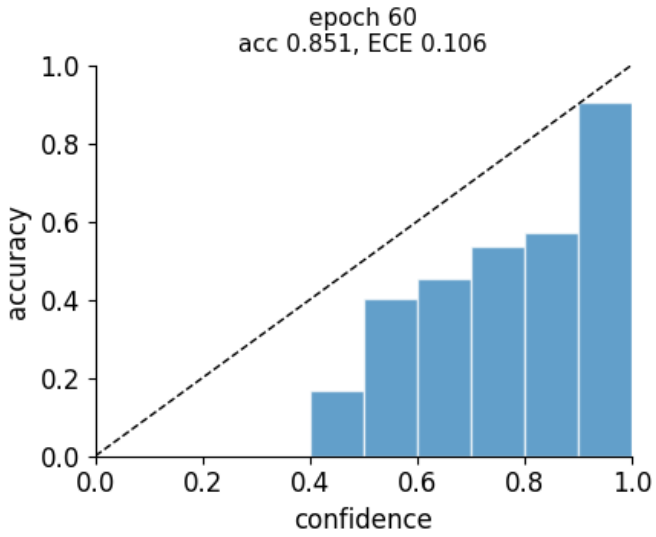
	accuracy at confidence c	100 inputs at confidence 0.8
calibrated	= c	80 correct
overconfident	< c	60 correct
underconfident	> c	95 correct

Calibration ≠ accuracy

Classifier (10 balanced classes)	Accuracy	Calibrated?
ignores the input, outputs 0.1 for every class	10%	yes
right on 99% of inputs, confidence 1.0 on all	99%	no, overconfident

Reliability diagrams and ECE

Bin test inputs by confidence: $\mathbf{B}_1$, …, $\mathbf{B}_m$. Plot accuracy against confidence per bin.

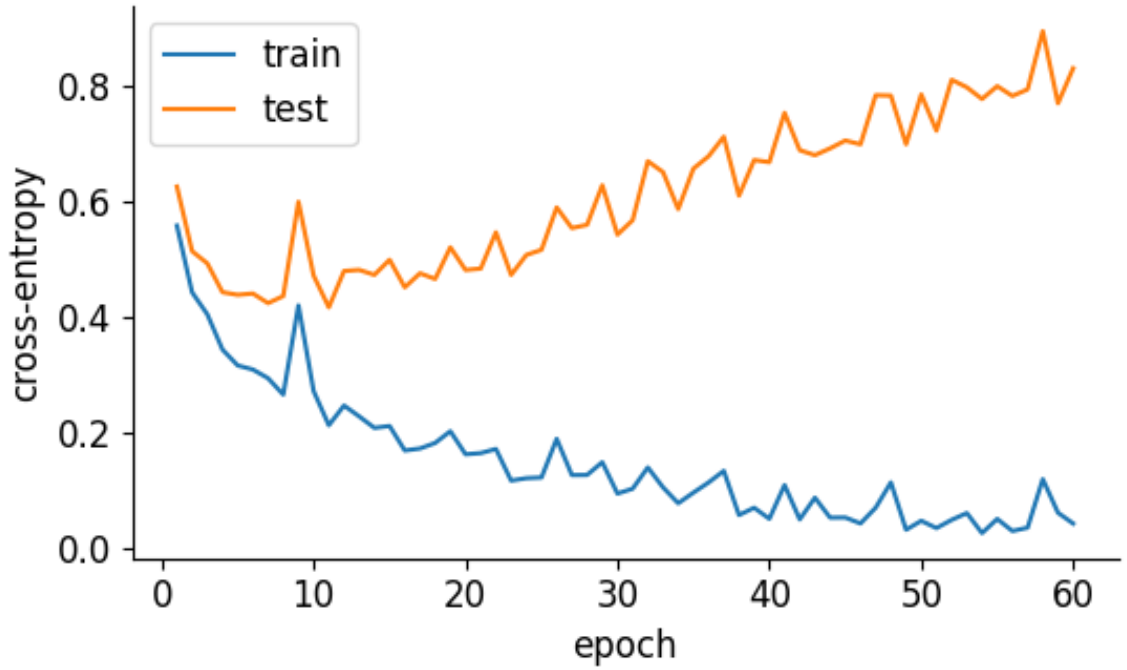


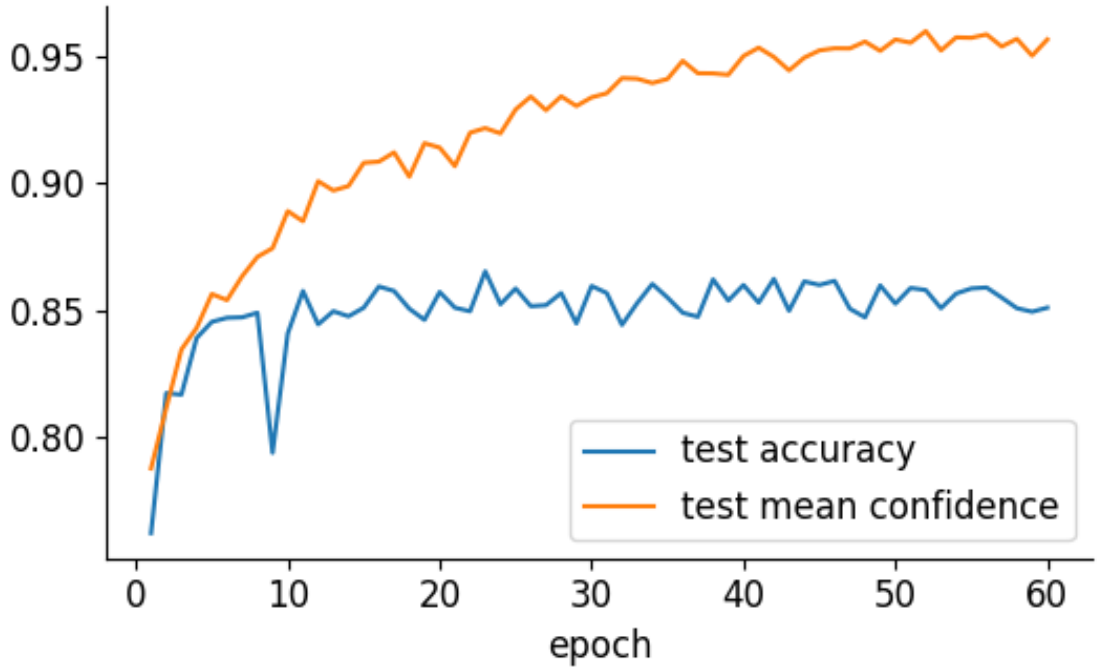
bars below the diagonal: overconfident (an MLP on FashionMNIST, Section 1.2)

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

depends on the number of bins (here 10); 0 for the useless classifier, so read it with accuracy

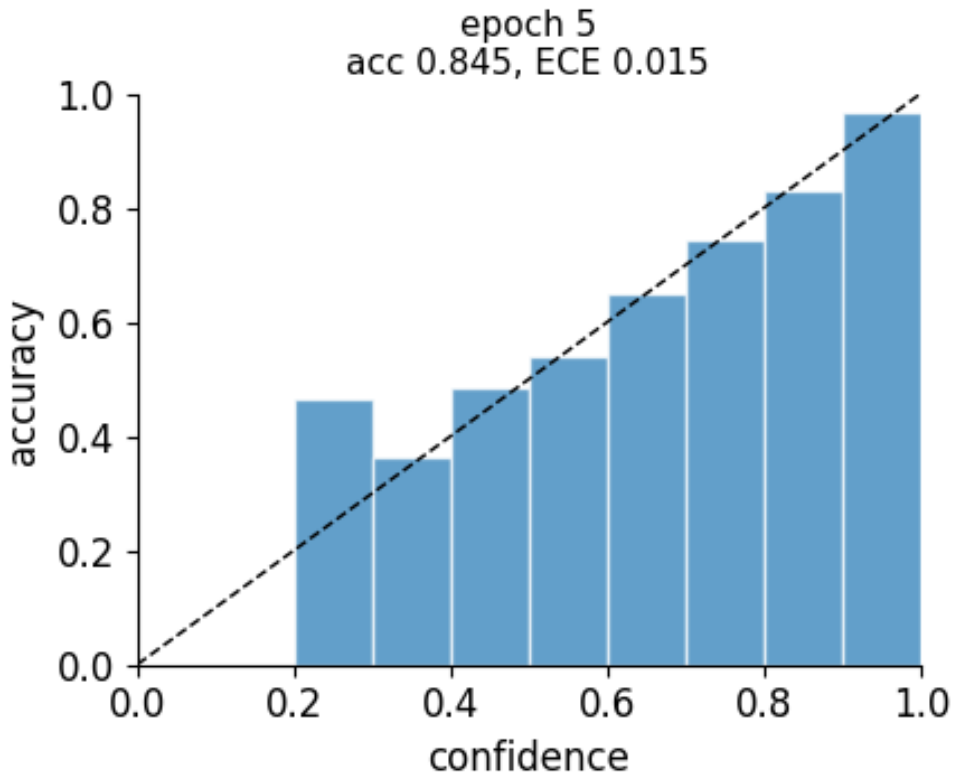
An overconfident network

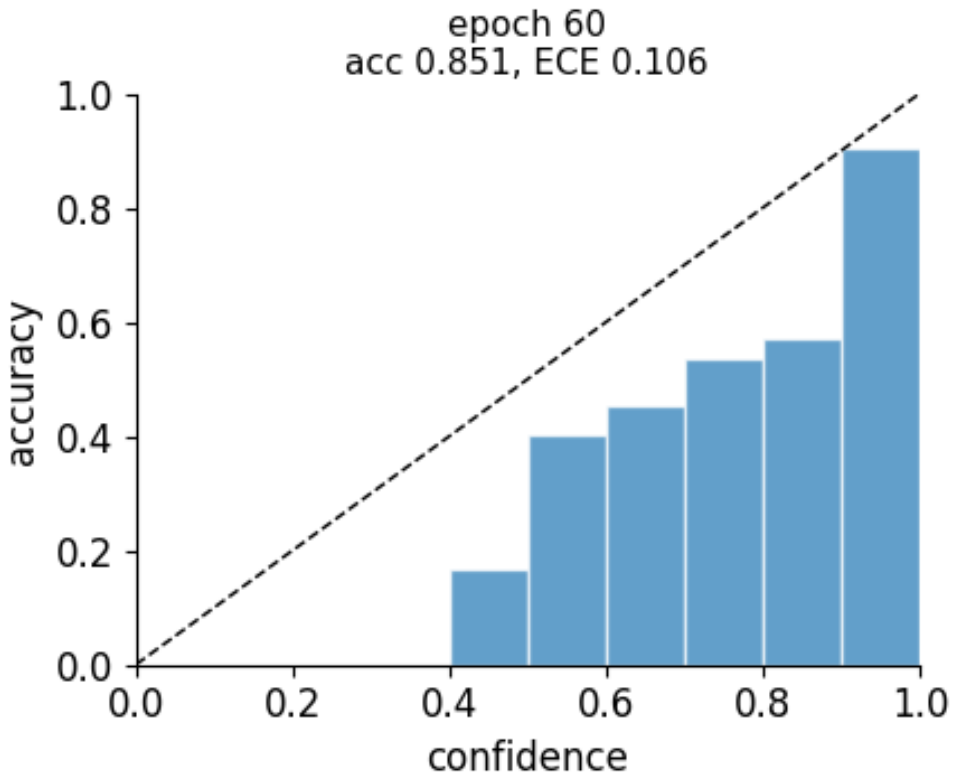




MLP on 10,000 FashionMNIST images, 60 epochs. Accuracy ≈ 0.85 throughout; mean confidence 0.86 → 0.96.

Reliability: epoch 5 vs epoch 60





	epoch 5	epoch 60
ECE	0.015	0.106

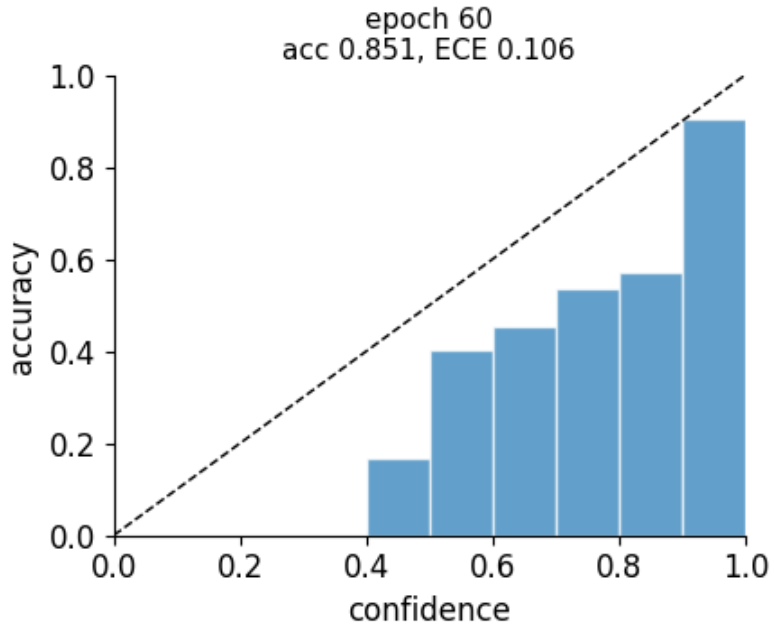
Overconfidence in modern networks

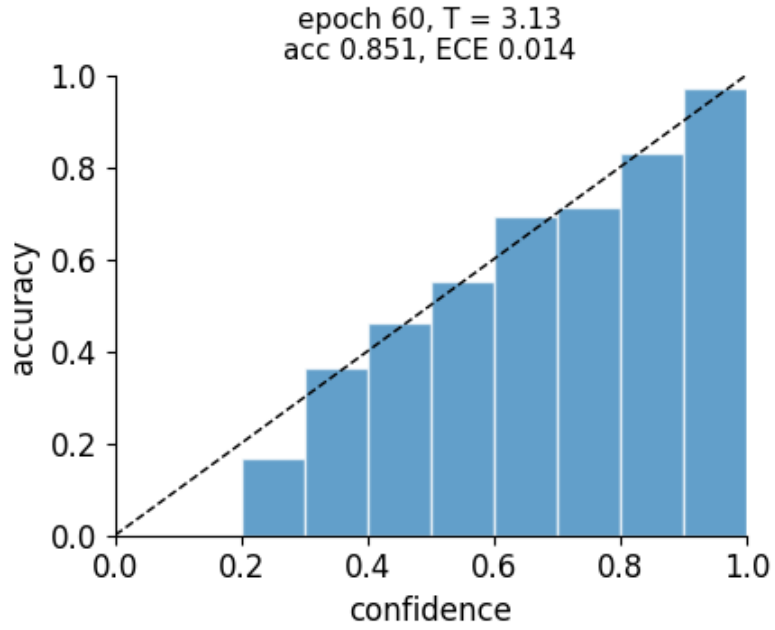
	CIFAR-100
LeNet (5 layers, 1998)	close to calibrated
ResNet-110	more accurate, markedly overconfident

- Depth, width, batch norm, less weight decay → more overconfident (Guo et al., 2017)
- Vision transformers are among the best calibrated (Minderer et al., 2021)

Temperature scaling

$$p^{(T)}(x) = \mathrm{softmax}(z(x) / T)$$





T = 3.13

fitted by minimizing cross-entropy on 5,000 validation images. ECE 0.106 → 0.014, same predictions.

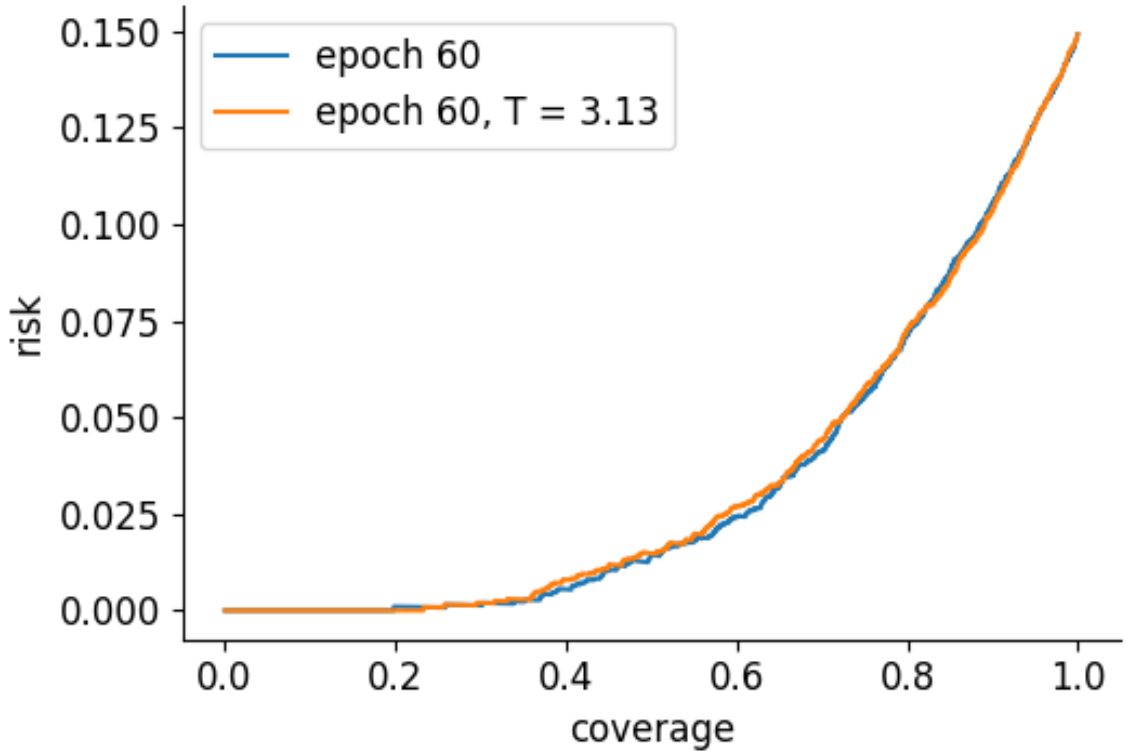
2

Abstention

Answer only when confident enough.

Selective prediction

$$\text{coverage} = P(\hat{p}(x) \geq \tau) \qquad \text{risk} = P(\hat{y}(x) \neq y \mid \hat{p}(x) \geq \tau)$$



- Sweep τ: the risk–coverage curve.
- Depends only on the ranking of the confidences, not on calibration.

Choosing a threshold

$$\widehat{\mathrm{risk}}(\tau) = 1 - \mathrm{mean}\left\{ \hat{p}(x) \ : \ \hat{p}(x) \geq \tau \right\}$$

	threshold τ	coverage at τ	risk at τ	predicted risk at τ
T = 1	0.993	0.747	0.055	0.000
T = 3.13	0.765	0.745	0.056	0.044

threshold: lowest τ with risk ≤ 5% on 5,000 validation images; measured on the test images

3

Distribution shift

Trained and evaluated on D, deployed on D’.

Distribution shift

Distribution shift: the distribution of inputs at deployment differs from the one the model was trained and evaluated on.

$$P_{D'}(x) \neq P_D(x)$$

different camera · different weather · different subjects

Shift vs adversarial perturbation

$$\text{adversarial}: \quad \mathbb{E}_{(x,\,y)\,\sim\,D}\left[\max_{\delta \in S} 1[f(x+\delta) \neq y]\right]$$

$$\text{shift}: \quad \mathbb{E}_{(x,\,y)\,\sim\,D'}[1[f(x) \neq y]]$$

worst case over a set chosen by an opponent vs average over a distribution nobody chose

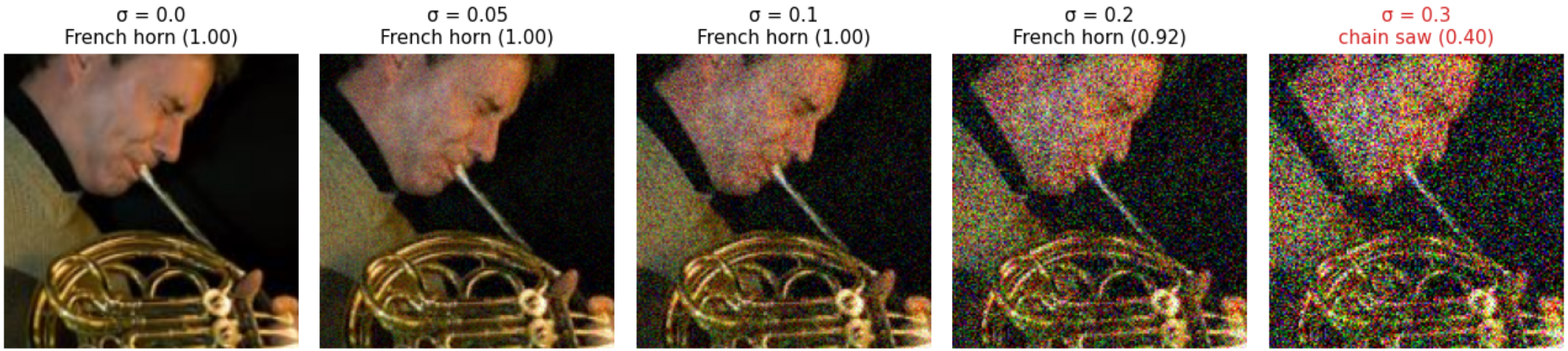
Accuracy and calibration are defined on a distribution

$$\mathrm{acc}_D = P_D(\hat{y}(x) = y) \qquad \mathrm{calibration}_D:\ P_D(\hat{y}(x) = y \mid \hat{p}(x) = c)\ \ \mathrm{vs}\ \ c$$

the shift error on the previous slide is 1 − acc_D’

- Under shift, each can change, and not necessarily together.
- p̂(x) is computed from the input alone: confidence can stay high while accuracy falls.

Gaussian noise on Imagenette



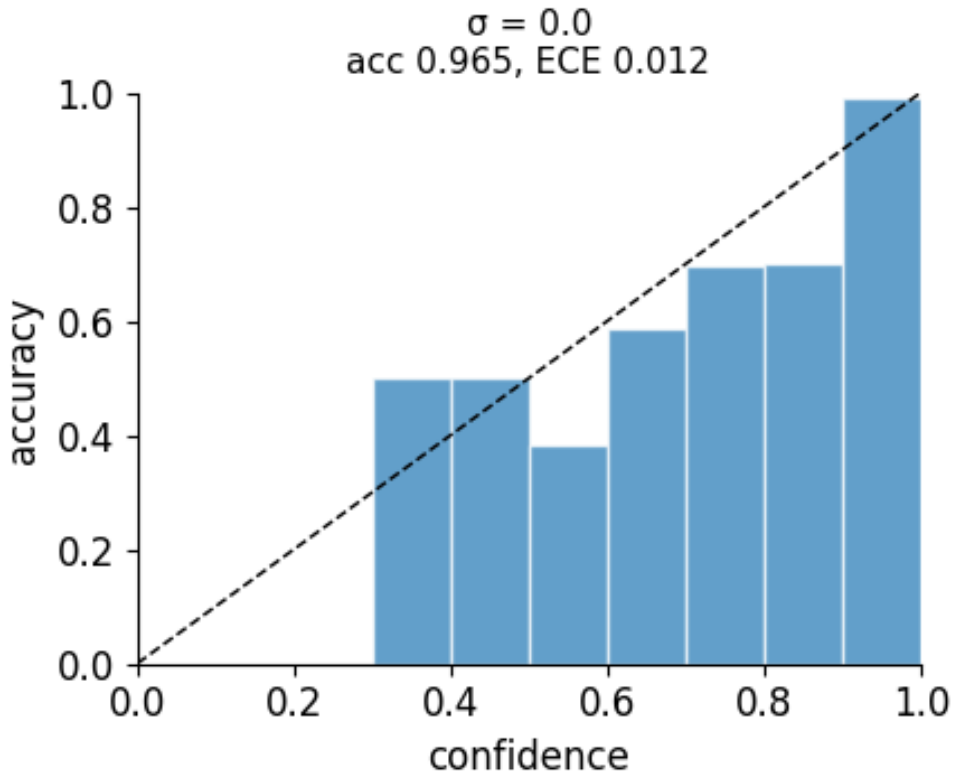
ResNet-18 (ImageNet) restricted to the 10 Imagenette classes. Title: prediction (confidence); red = wrong. Noise as in ImageNet-C (Hendrycks & Dietterich, 2019).

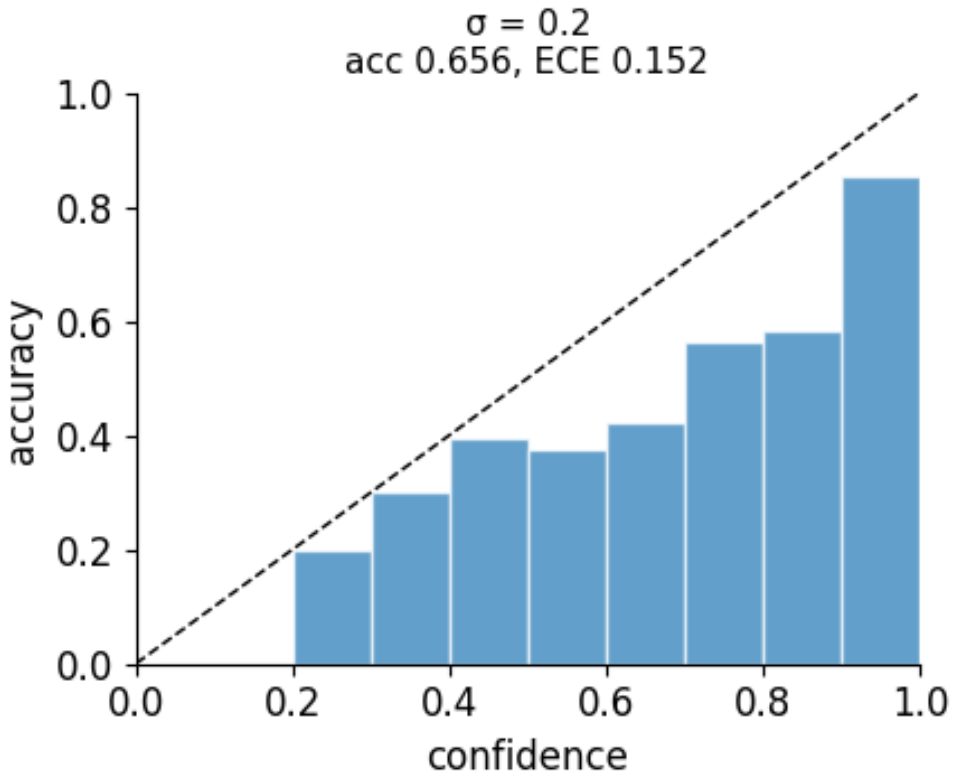
Accuracy falls faster than confidence

σ	accuracy	mean conf	ECE	ECE, T = 1.32
0	0.965	0.975	0.012	0.005
0.05	0.943	0.964	0.021	0.010
0.1	0.889	0.931	0.042	0.018
0.2	0.656	0.808	0.152	0.093
0.3	0.316	0.779	0.463	0.386

T fitted on clean photographs.

Reliability under noise





Thresholds under shift

σ	coverage at τ	risk at τ	predicted risk at τ
0	0.934	0.008	0.003
0.05	0.899	0.016	0.005
0.1	0.809	0.035	0.008
0.2	0.502	0.147	0.019
0.3	0.428	0.596	0.030

τ: lowest threshold with risk ≤ 1% on clean photographs

4

Language models: multiple choice

The answer is one letter, so Section 1 applies directly.

From classes to token sequences

classifier $$p(y \mid x), \quad y \in \{1, \ldots, K\}$$

language model $$p(s \mid x) = \prod_{t=1}^{|s|} p(s_t \mid x, s_{<t}), \quad s \in V^*$$

V*: all finite sequences of tokens

- No fixed answer set: the max over all sequences cannot be taken.
- One claim, many sequences: “Paris”, “It is Paris.”, “The capital is Paris.”
- A sequence's probability shrinks with its length.

Multiple choice (Section 4) returns to a fixed set of letters; free-form answers (Section 5) need new measures.

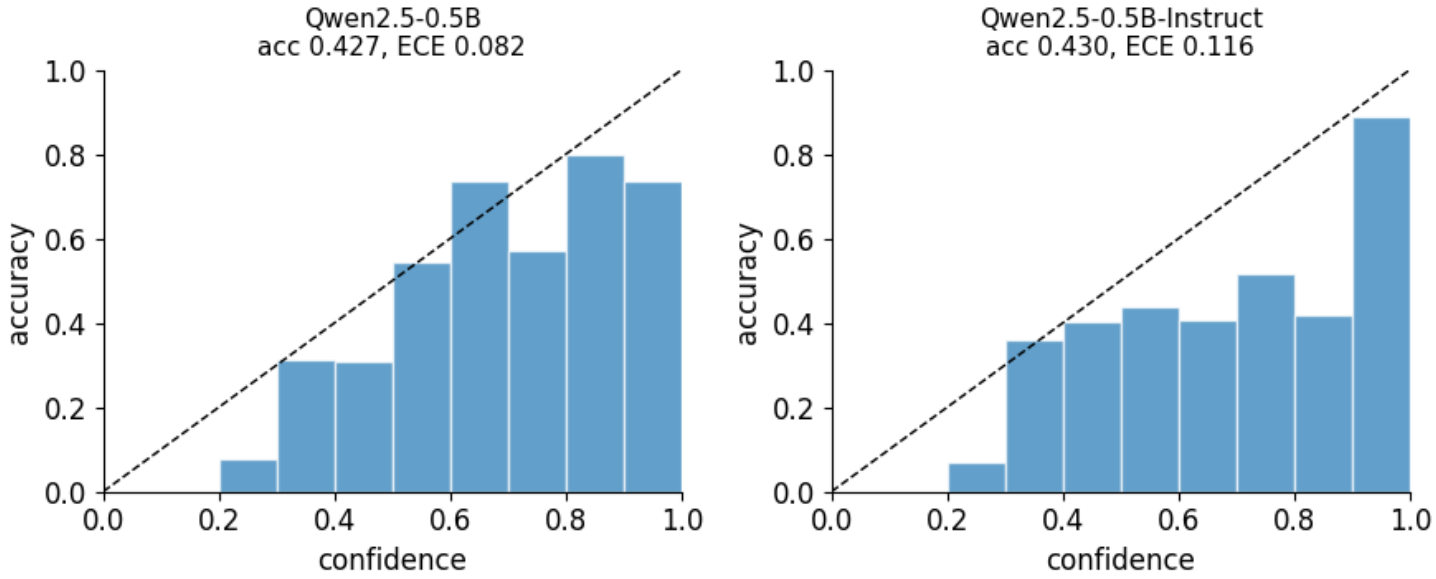
Confidence from letter probabilities

$$\hat{p}(x) = \max_{L} \frac{p(L|x)}{\sum_{L' \in \{A,B,C,D\}} p(L' \mid x)}$$

```
Question: …
A. …   B. …   C. …   D. …
Answer:  ▸ next token ∈ { A, B, C, D }
```

400 MMLU questions (Hendrycks et al., 2021); format of Kadavath et al. (2022). Qwen2.5-0.5B base vs Instruct.

Base vs post-trained



	base	Instruct
accuracy	0.427	0.430
ECE	0.082	0.116

ECE(Instruct) − ECE(base) = **0.034**

95% bootstrap interval: **[−0.02, 0.08]**

Resample the 400 questions with replacement and recompute the difference. The interval includes 0, so the difference is not significant.

Post-training and calibration

$$\text{pretraining}: \min \mathbb{E}[-\log p(\text{text})] \qquad \text{post}-\text{training}: \max_{\pi} \mathbb{E}_{y\sim\pi(\cdot|x)}\left[r(x,y)\right]$$

- Reward depends on the response, not on the probability the model gave it.
- GPT-4: “the pre-trained model is highly calibrated” … “the post-training hurts calibration significantly” (OpenAI, 2023)
- RLHF models “naively seem miscalibrated”, largely fixed by a temperature (Kadavath et al., 2022)

5

Language models: free-form answers

No single probability: the same claim has many token sequences.

Four measures

	Measure	Uses
5.2	verbalized confidence	a number the model writes
5.2	P(True)	probability of one token, judging its own answer
5.3	predictive entropy	probabilities of sampled answers
5.4	semantic entropy	probabilities of sampled meanings

The questions and answers

```
Answer with just the answer, in as few words as possible.
Question: What is the capital of Australia?
```

Qwen2.5-0.5B-Instruct, 80 questions written for this lecture: capitals, authors, painters, composers, chemistry, dates, geography

Per question	How	Used for
1 answer	greedy: most likely token at each step	graded: correct if it contains an accepted answer
10 samples	temperature 0.7, each with log p(s | x)	not graded; inputs to the entropies in 5.3 and 5.4

55 of 80

greedy answers correct

Each measure gives an uncertainty for each question. All four are judged against the same label: whether the greedy answer is correct.

Asking the model

Verbalized confidence

```
Q: What is the capital of Australia?
A: Sydney
How confident are you that your
answer is correct? Reply with only
a number from 0 to 100.
```

confidence = stated number / 100

P(True)

```
Question: What is the capital of
Australia?
Proposed answer: Sydney
Is the proposed answer correct?
Reply True or False.
```

confidence = P(True | question, answer)

Both judge the model's own greedy answer. Uncertainty u = 1 − confidence.

Detecting wrong answers

$$\mathrm{AUROC} = P(u(x_{\mathrm{wrong}}) > u(x_{\mathrm{correct}}))$$

x_{wrong}, $x_{correct}$: a random question whose greedy answer is wrong / correct

u(x): the uncertainty a measure gives for question x

1: every wrong answer more uncertain than every correct one **0.5**: unrelated to correctness

	AUROC
verbalized confidence (states 95 for 76 of the 80 answers)	0.54
P(True)	0.77

Token-level uncertainty

$$H(x) = -\sum_{s} p(s \mid x) \log p(s \mid x) \;=\; \mathbb{E}_{s \sim p(\cdot \mid x)}[-\log p(s \mid x)]$$

$$\approx \; -\frac{1}{M} \sum_{i=1}^{M} \log p(s^{(i)} \mid x), \qquad s^{(1)}, \ldots, s^{(M)} \sim p(\cdot \mid x)$$

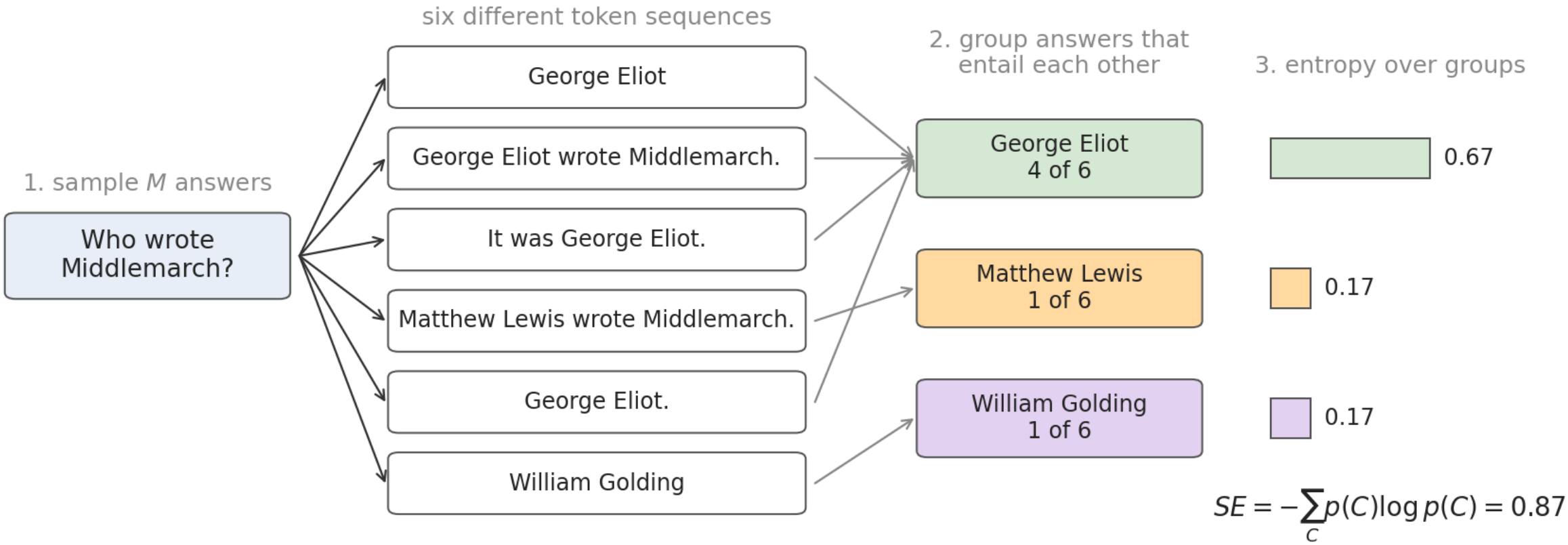
length-normalized $$-\frac{1}{M} \sum_{i=1}^{M} \frac{1}{|s^{(i)}|} \log p(s^{(i)} \mid x)$$

M = 10 samples per question. Low when one reply has most of the probability, high when it is spread. The samples are not graded.

“Paris”, “Paris.”, “The capital of France is Paris.” are three sequences with one meaning

Semantic entropy

$$SE(x) = -\sum_C p(C \mid x) \log p(C \mid x) \qquad p(C \mid x) = \sum_{s \in C} p(s \mid x)$$



Same meaning = bidirectional entailment, judged by an NLI classifier (Kuhn et al., 2023; Farquhar et al., 2024).

Examples

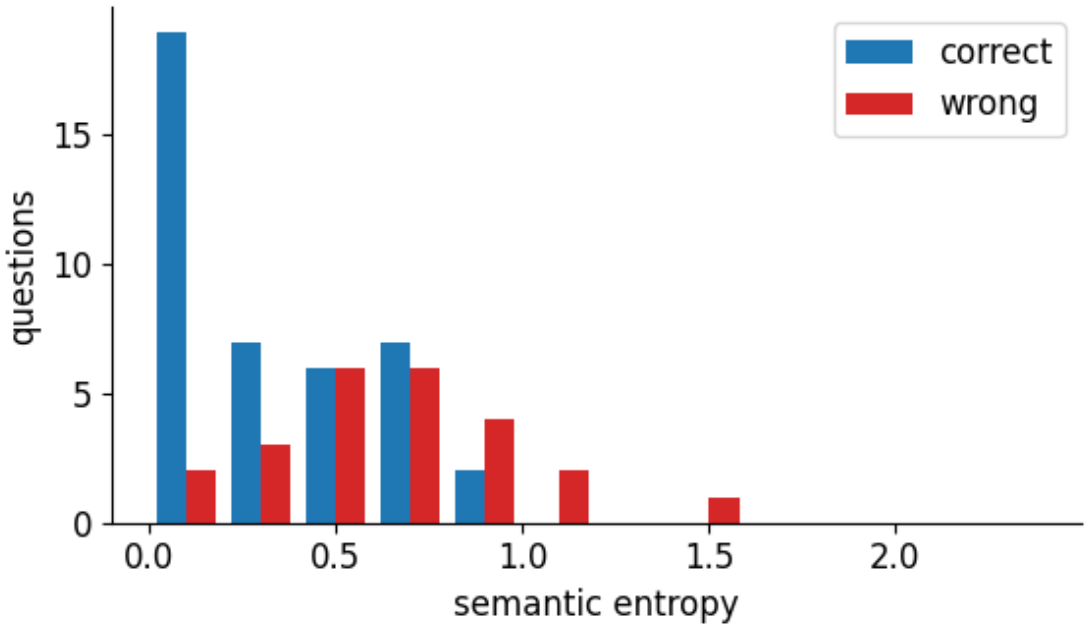
```
Battle of Hastings  [correct]   SE 0.36
   (6) 1066 | 1066.
   (3) The Battle of Hastings took place in 1066 AD.
   (1) The Battle of Hastings took place in 1066 when King Harold II

Who wrote The Tale of Genji?  [wrong]   SE 1.09
   10 clusters of 1: Kunio Amano | Koichiro Genda | Tosuke Akimitsu | …

Capital of Australia  [wrong]   SE 0.30
   (9) Sydney | Sydney is the capital of Australia.
   (1) Canberra
```

Comparing the measures

Measure	AUROC
verbalized confidence	0.54
P(True)	0.77
predictive entropy	0.75
length-normalized predictive entropy	0.80
discrete semantic entropy	0.79
semantic entropy	0.82

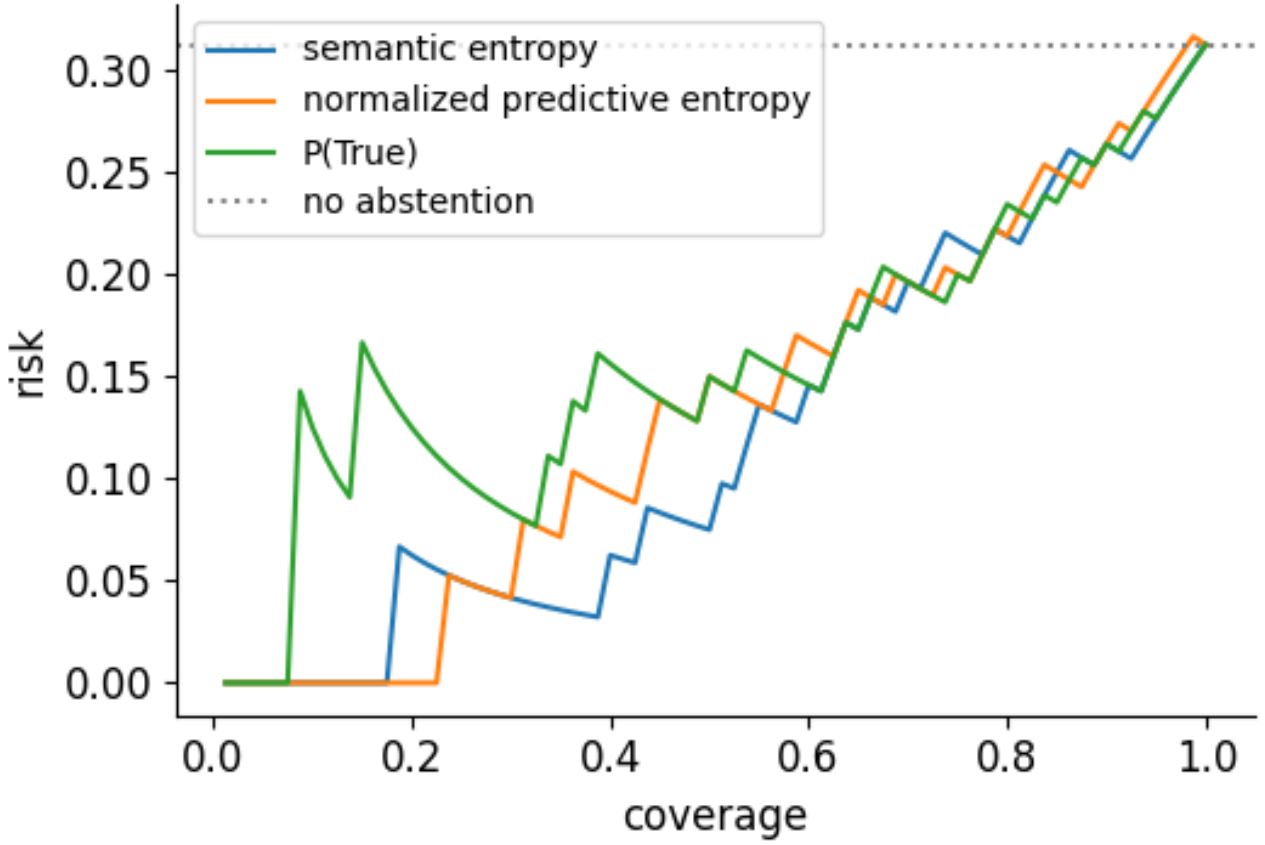


Consistent errors

Question	Answer	Same meaning
What is the capital of Australia?	Sydney	9 / 10
What is the capital of Peru?	Cusco	7 / 10
Who painted The Persistence of Memory?	Vincent van Gogh	10 / 10
What is the chemical symbol for tungsten?	Wt	7 / 10
Who discovered the electron?	Rutherford	7 / 10
What language is mainly spoken in Brazil?	Spanish	8 / 10

no disagreement among samples to measure; Farquhar et al. target confabulations, answers that change when asked again

Abstaining on uncertain answers



- no abstention: risk 0.31
- semantic entropy: risk < 0.1 up to 40% coverage, 0 below ≈ 18%
- no measure reaches 0 risk at higher coverage: the consistent errors

Summary

- **Calibration**: accuracy at confidence c equals c. Cross-entropy training can break it; temperature scaling repairs it in distribution.
- **Selective prediction**: the curve needs a good ranking; predicting the risk needs calibration.
- **Distribution shift**: accuracy fell faster than confidence; clean-data temperatures and thresholds failed.
- **Multiple choice**: letter probabilities are a confidence; post-training tends to make them overconfident.
- **Free-form**: semantic entropy did best; no sampling measure catches errors the model repeats.

Reading

Required

Guo et al., On Calibration of Modern Neural Networks (arXiv 1706.04599)

Farquhar et al., Detecting Hallucinations in Large Language Models Using Semantic Entropy (Nature, 2024)

Optional

Ovadia et al., Can You Trust Your Model's Uncertainty? · Kadavath et al., Language Models (Mostly) Know What They Know · Kuhn, Gal & Farquhar, Semantic Uncertainty · Kalai et al., Why Language Models Hallucinate

Next lecture: agents, and the threats that come with them.

AI (Claude Code) was used to help create this deck.