Calibrated Large Language Models for Binary Question Answering

Patrizio Giovannotti Alex Gammerman

Department of Computer Science
Royal Holloway University of London

13th Symposium
on Conformal and Probabilistic Prediction with Applications

Language Models and Uncertainty

Just to be really frank, the thing that's really missing today is that a model doesn't raise its hands and say "Hey, I'm not sure, I need help"

— Vik Singh, Head of Microsoft Copilot

- Quantifying uncertainty for a sentence is challenging
- Easier case: binary answers ("Yes" / "No")
- **Calibrated** models can offer a reliable estimate of their uncertainty

Calibration

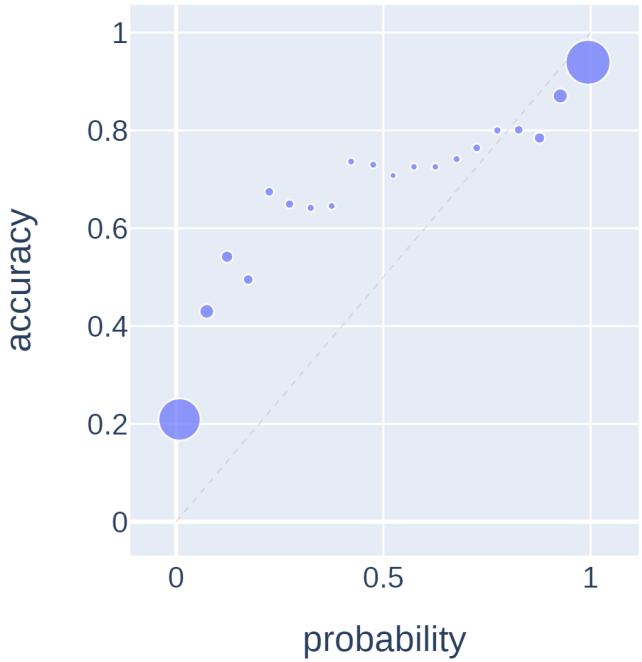
Given a prediction $\hat{Y}$ for the label Y, returned with an estimated confidence $\hat{P}$, an ML model is *perfectly calibrated* if

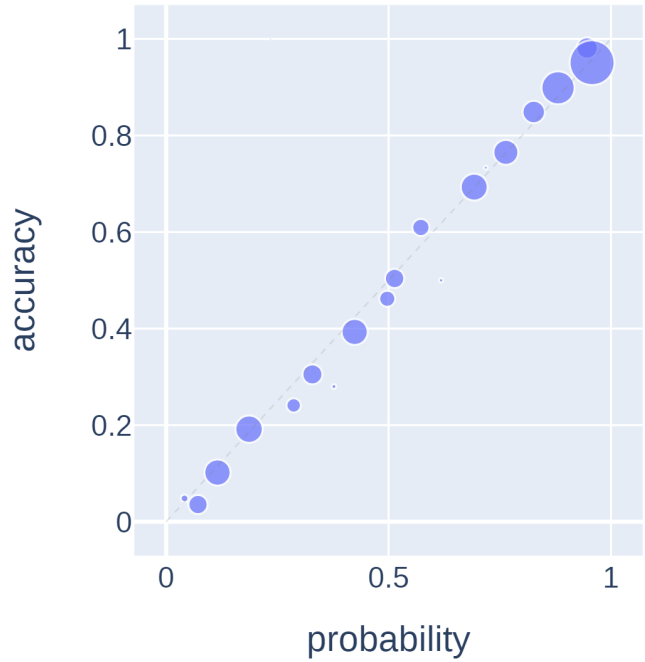
$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p, \qquad \forall p \in [0, 1]$$

Calibration Techniques

- Isotonic regression
- Platt's scaling
- Histogram binning
- **Temperature scaling**
- **Venn–Abers predictors**

Reliability Bubble Chart





Background

Language Model task (1975)

Compute the probability of a given sequence of words

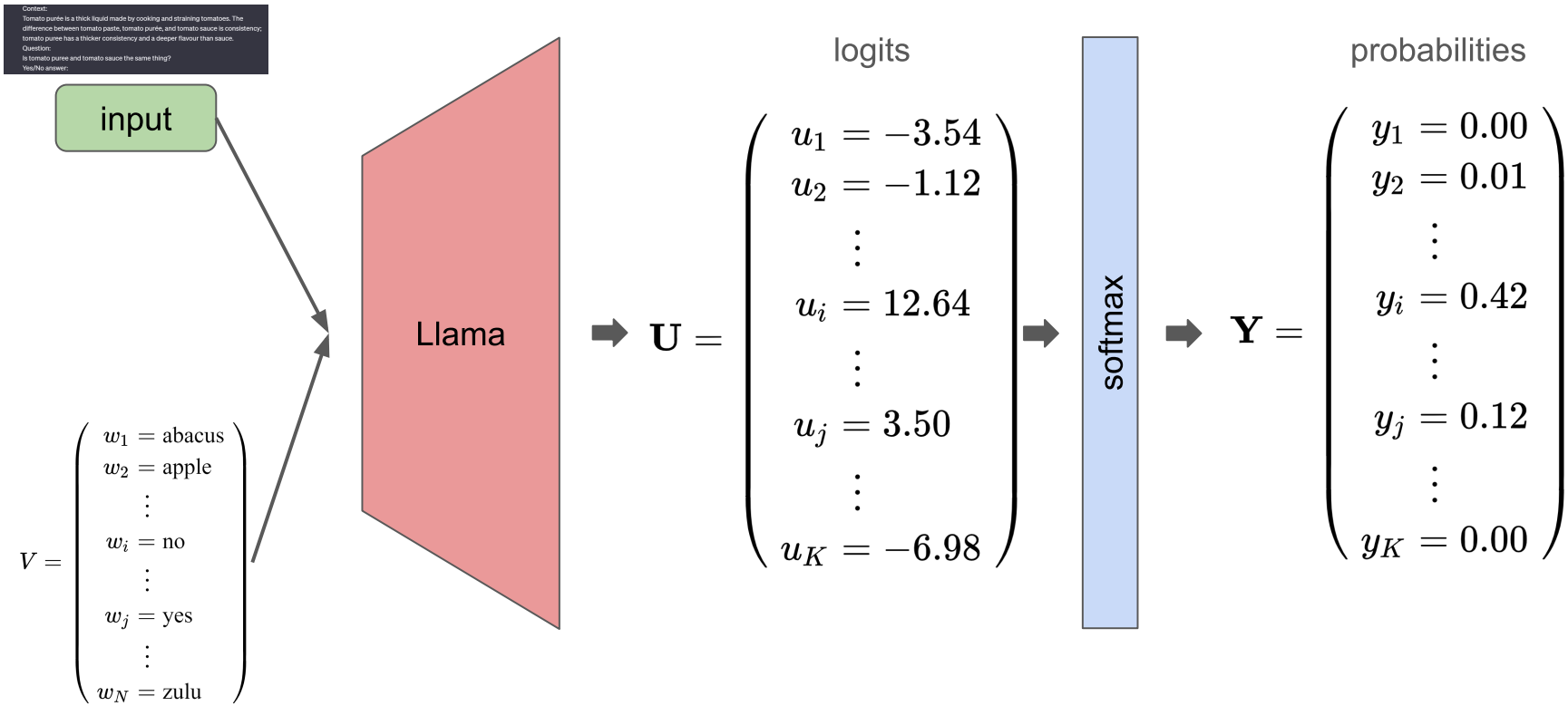
$$P(w_{1:n}) = P(w_1, w_2, \ldots, w_n),$$

$w_i \in W$ (the vocabulary) $\forall i = 1, \ldots, n$.

This relies on computing the probability of each word w_i given the previous words:

$$P(w_{1:n}) = \prod_{i=1}^{n} P(w_i \mid w_{1:i-1})$$

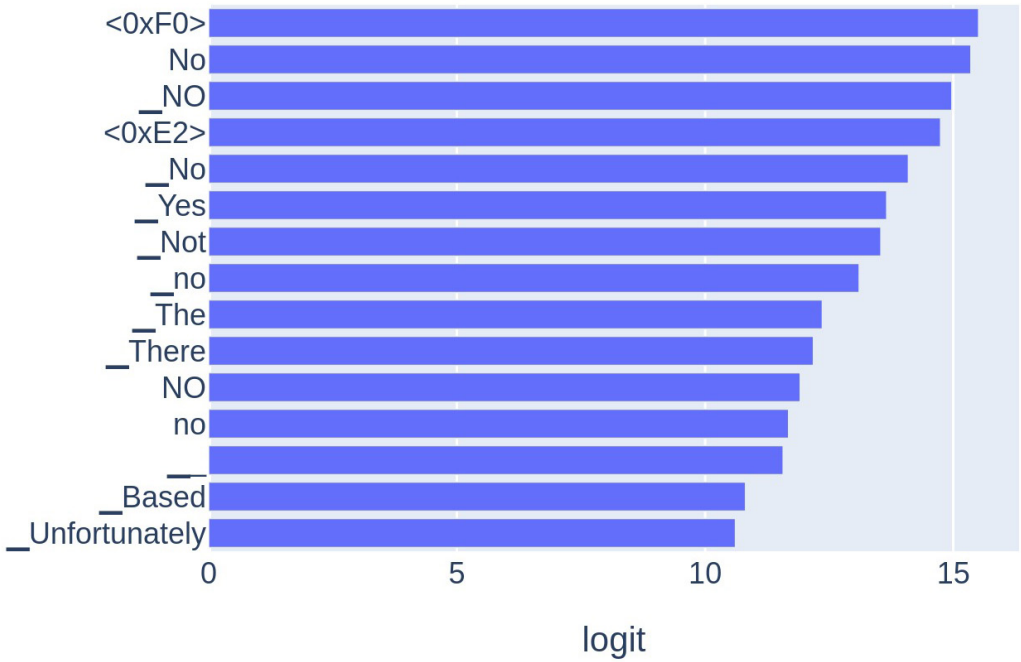
Large Language Models



Binary Question Answering

Example

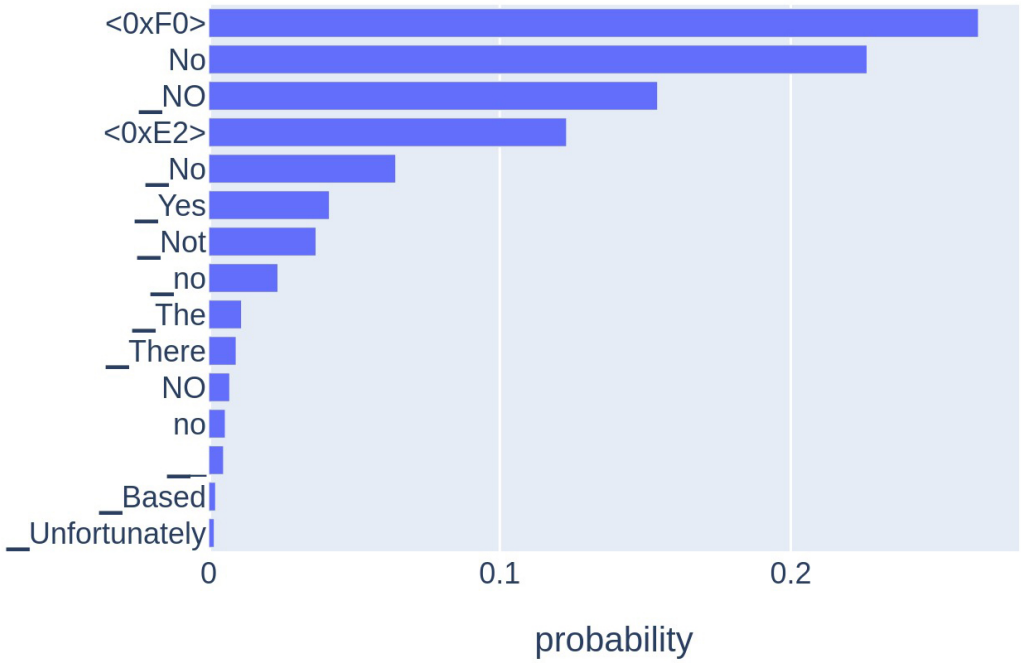
Context: "The show was renewed on April 4, 2014, for a sixth and final season." Question: "Is there a season 7 of *Republic of Doyle*?" Yes or No? Answer:



Binary Question Answering

Example

Context: "The show was renewed on April 4, 2014, for a sixth and final season." Question: "Is there a season 7 of *Republic of Doyle*?" Yes or No? Answer:



Temperature Scaling

Find $\tau \in \mathbb{R}$ s.t. the scores of each word k

$$\text{softmax}_\tau(u_k) = \frac{\exp(u_i/\tau)}{\sum_{k=1}^{K} \exp(u_k/\tau)}$$

minimise the cross-entropy loss over a separate validation set.

Temperature Effects

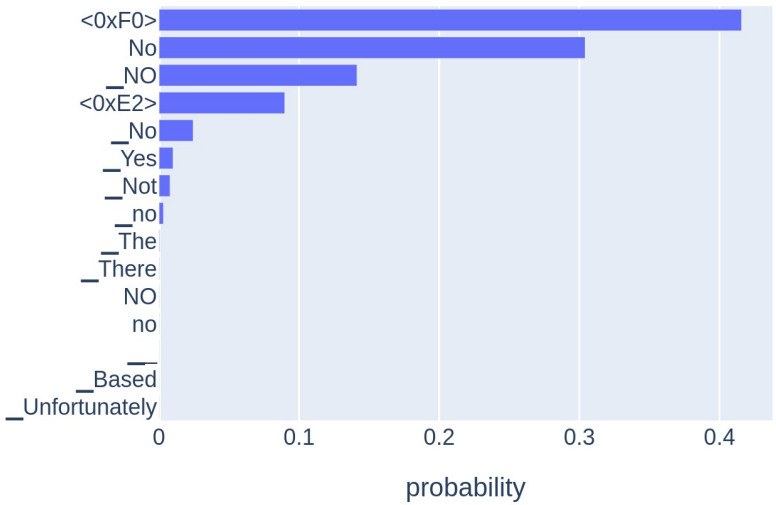


Figure 1: Temperature = 0.5

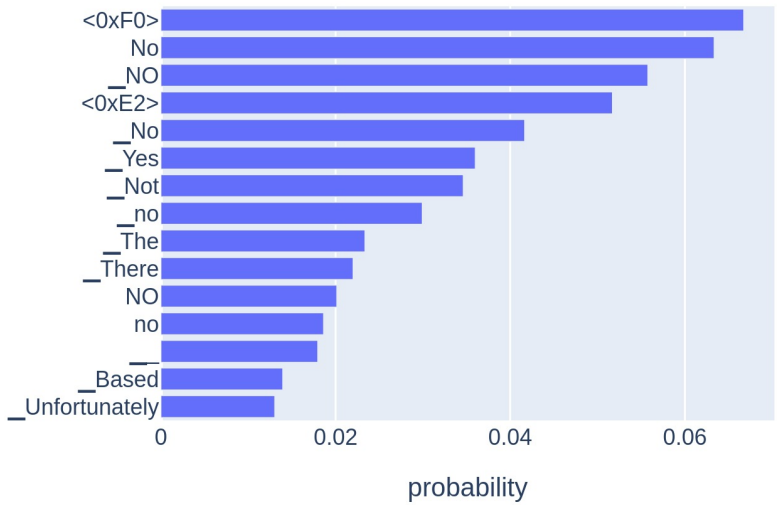
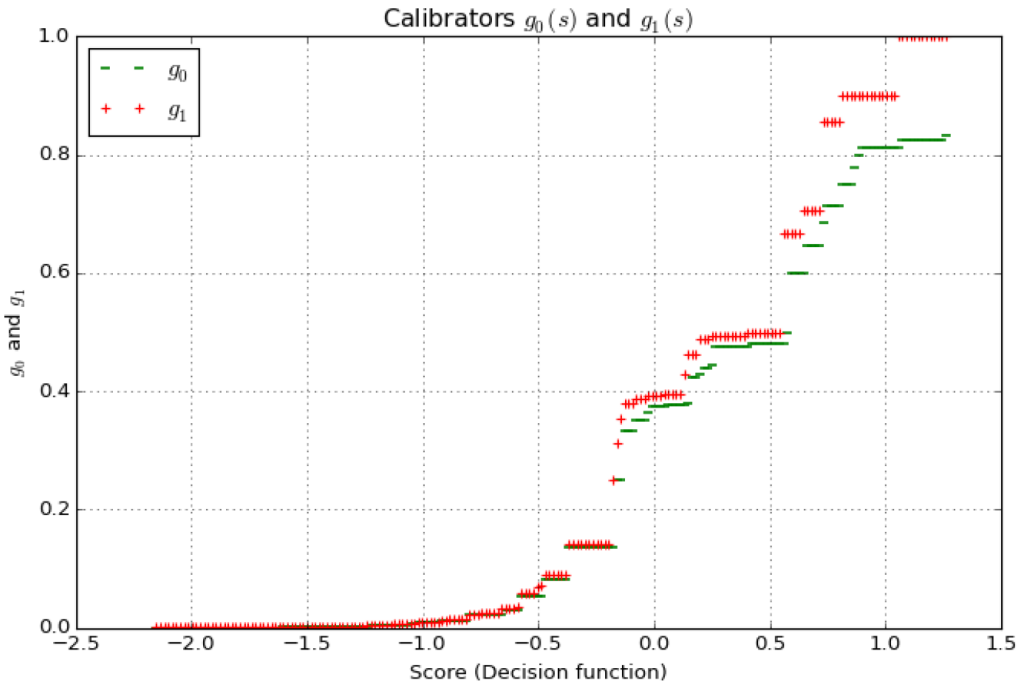


Figure 2: Temperature = 3

Inductive Venn–Abers Predictors



- Nonparametric approach to calibration
- For each prediction, output:
  - Prediction interval $[p_0, p_1]$
  - Point estimate $p_{\text{single}} = \frac{p_1}{1-p_0+p_1}$
- Zero-shot setting: no need for proper training set

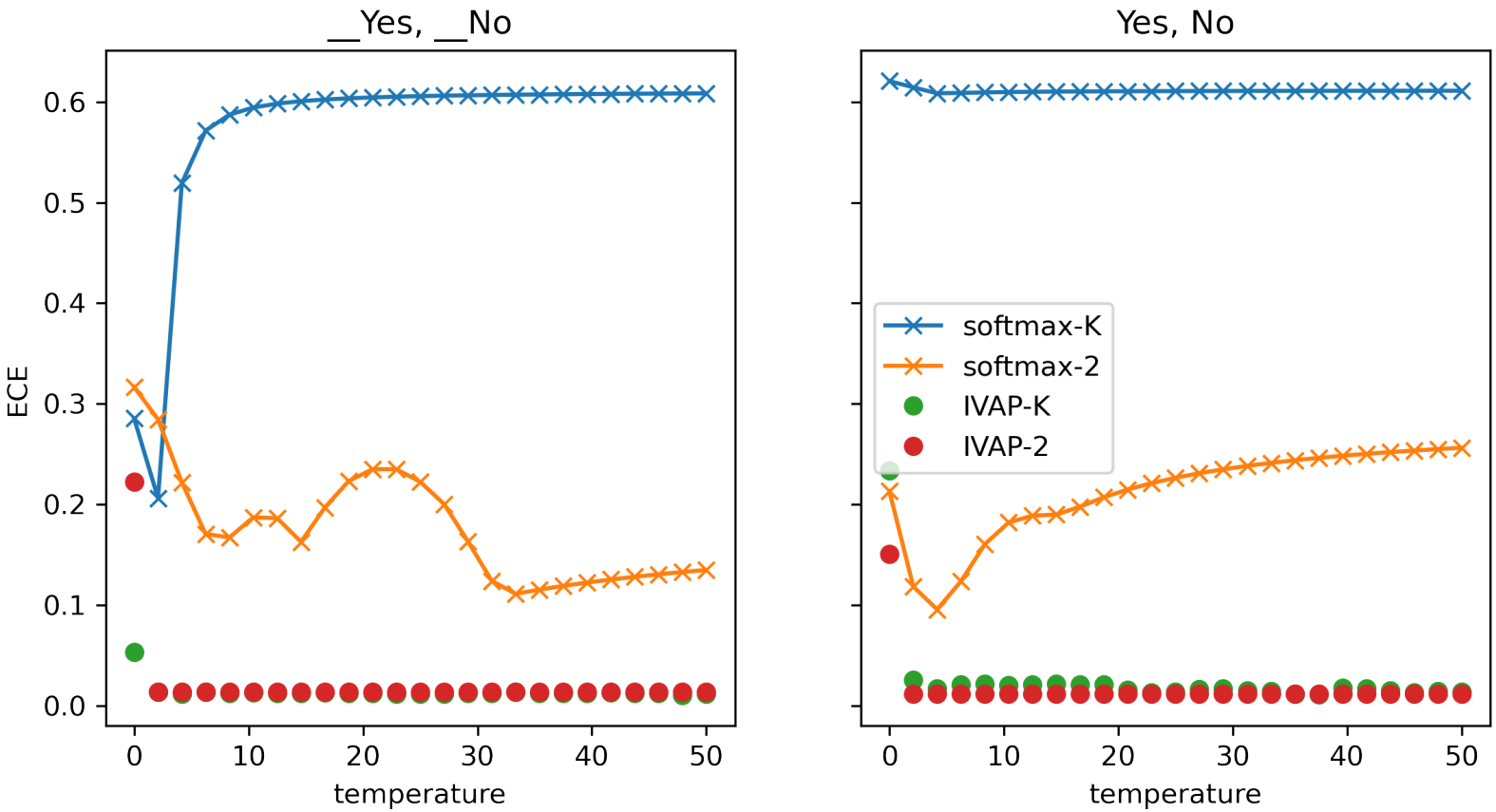
Setup and Evaluation

Token choices

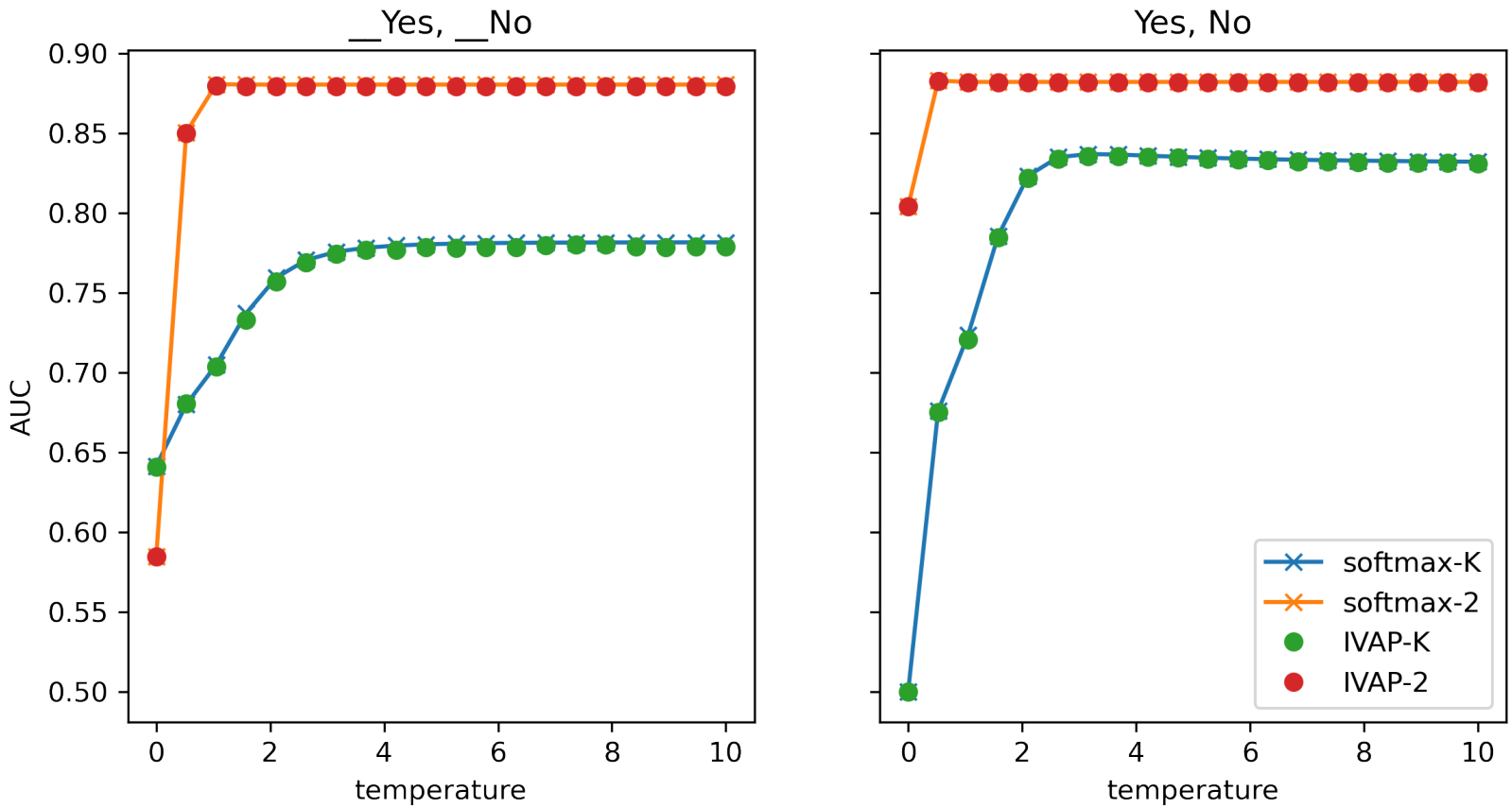
softmax-K	"Yes" and "No" scores selected from the softmax over all K logits
softmax-2	Scores from softmax computed over the sole "Yes" and "No" logits
IVAP-K	Calibrated version of 1, via the inductive Venn–Abers predictor
IVAP-2	Calibrated version of 2, via the inductive Venn–Abers predictor

- Over a range $[\tau_0, \tau_1]$ of temperature choices.
- Evaluate with Expected Calibration Error (ECE) and AUC

Results – Calibration



Results – Prediction Quality



Examples

High confidence

Context:
"New Balance maintains a manufacturing presence in the United States, as well as in the United Kingdom for the European market, where they produce some of their most popular models such as the 990 model [...] They are the second most-renown American sporting company, after Nike."
Question: "Are New Balance and Nike the same company?" Yes or No?
Answer:

$$p_0(\texttt{Yes}) = 0.00 \quad , \quad p_1(\texttt{Yes}) = 0.04$$

$$p(\texttt{Yes}) = \mathbf{0.038}$$

Examples

Low confidence

Context:
"System of a Down, [...], is an Armenian-American heavy metal band from Glendale, California, formed in 1994. The band currently consists of Serj Tankian (lead vocals, keyboards), Daron Malakian (vocals, guitar), Shavo Odadjian (bass, backing vocals) and John Dolmayan (drums)."
Question: "Does system of a down have 2 singers?" Yes or No?
Answer:

$$p_0(\texttt{Yes}) = 0.493 \quad , \quad p_1(\texttt{Yes}) = 0.507$$

$$p(\texttt{Yes}) = \mathbf{0.5}$$

Conclusion

- We presented a competitive method to calibrate the output of LLMs in the binary QA setting
  - no further training of the LLM required
  - no special assumption on the data distribution
- Venn–Abers outperforms temperature scaling, invariant w.r.t.
  - temperature
  - tokens (words) chosen to determine the answer
- First step towards a reliable and safer AI, where models can precisely determine and communicate their degree of uncertainty

Next Steps

- Case with n. labels > 2
- *White-box* approach (no access to scores required)
- Sentence-level uncertainty