Combining Confidence Elicitation and Sample-based Methods for Uncertainty Quantification in Misinformation Mitigation

Mauricio Rivera[1], Jean-François Godbout[2], Reihaneh Rabbany[1], Kellin Pelrine[13]

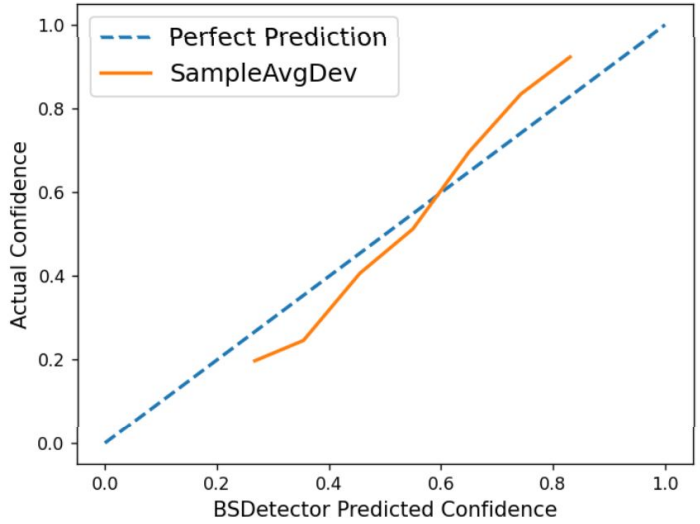
[1]McGill University; Mila [2]Université de Montréal; Mila [3]Stitch

Best sample-based consistency method: prompt for score 0-100, average over multiple runs of deviation from 50

Method	α	ECE	Brier Score
self-consistency	0.4	0.119	0.324
selfcheckGPT	0.7	0.119	0.330
PredClassMargin	0.4	0.131	**0.316**
SampleAvgDev	0.9	**0.076**	0.334
Norm. std	0.8	0.112	0.322
Deviation-Sum	0.6	0.133	0.321

Best verbalized confidence method: 2-step - predict first, then prompt again for uncertainty

Combine them: implement BSDetector framework[1] in this domain



[1]Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness. Chen & Mueller 2023

Uncertainty Resolution in Misinformation Detection

Yury Orlovskiy[1], Camille Thibault[2], Anne Imouza[3], Jean-François Godbout[2], Reihaneh Rabbany[3], Kellin Pelrine[34]

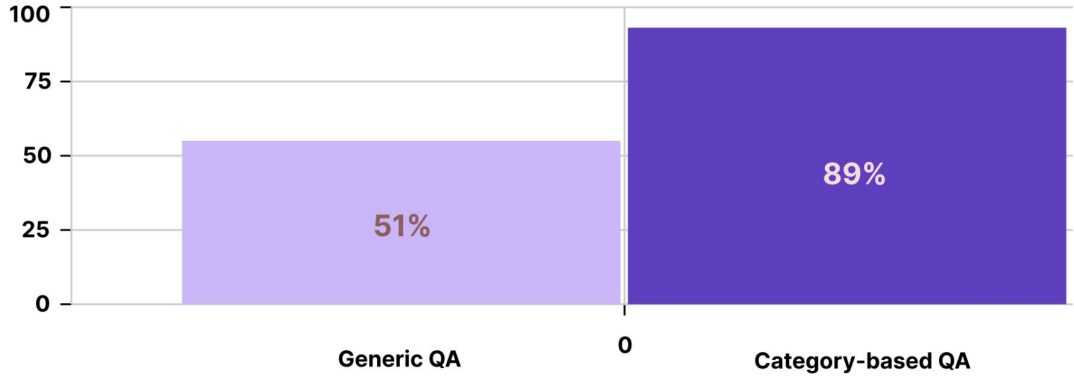
[1]UC Berkeley [2]U de Montréal [3]McGill U; Mila [4]Stitch

Categorize missing info (e.g., person, location, date, …)

Query user generically or with category-based prompt

Increase:

- **Answerability**
- **Resolved Questions**
- **Performance**



Experiment	Macro F1 (%)	Accuracy (%)	Percent Resolution (%)
Baseline (uncertainty disabled)	56.54	79.44	93.49
Baseline (uncertainty enabled)	71.76	91.28	16.70
Fill-in-the-blank method	79.60	91.79	20.09
Category-based QA	85.43	91.03	22.72
Category-based QA (uncertainty disabled)	68.90	81.10	90.30
Oracle Benchmark	96.71	99.16	69.41