LLM Governance Elements: Detection and Alignment

Prasanna Sattigeri
Principal Research Scientist
Trustworthy Machine Intelligence
IBM Research AI

21 June 2024

What does it take to trust an LLM based system?

Some risks are the **same as in traditional data science**

- poor predictive accuracy
- lack of fairness and equity
- lack of explainability
- model uncertainty
- distribution shifts
- poisoning attacks
- evasion attacks
- extraction attacks
- inference attacks
- model transparency

Occur when LLMs are used in "classical ML" tasks, e.g., prediction and classification, and have well-defined metrics and defenses

But many risks are **entirely new in generative settings**

- hallucinations
- lack of factuality or faithfulness
- lack of source attribution
- privacy leakage
- toxicity, profanities, and hate speech
- bullying and gaslighting
- prompt injection attacks
- social bias

Occur when LLMs are used in generative tasks, and do not yet have well-defined metrics and defenses.

Detectors for improving trust throughout LLM Lifecycle

Detector: a component that analyzes data, model input and/or model output to assess the likelihood of a particular AI risk

Data Curation for Training
Detects training data issues

- Runs offline, prior to deployment
- *Significant throughput requirements*

Model Risk Assessment
Comprehensive quantitative benchmarking for various AI risks

- Runs offline, prior to deployment
- *Liberal latency and medium throughput requirements*

Model Alignment
Improve model to satisfy enterprise principles

- Runs offline, prior to deployment
- *Liberal latency and medium throughput requirements*

Deployment

"Guardrails"
Online risk detection

- Runs during prompt/generation interaction
- *Significant latency requirements*

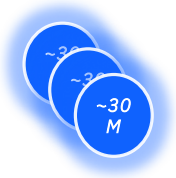
Observability/Monitoring
Online batch risk evaluation

- Runs periodically online on logged payload generations
- *Medium latency and high throughput requirements*

Detectors for Safe and Reliable LLMs: Implementations, Uses, and Limitations

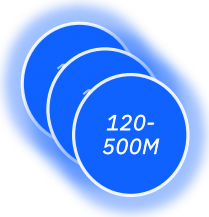
9 Mar 2024 · Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, Marcel Zalmanovici ·

cost/latency/accuracy/throughput tradeoffs



Many tiny models tuned for throughput

High Throughput
Low Latency
Low Cost



Many small models tuned for different tasks

Medium Throughput
Low Latency
Low Cost

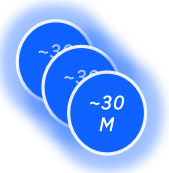
Single medium LM Tuned for all trust workloads

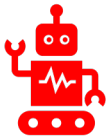
Medium Throughput
Medium Latency
Medium Cost

Single General LLM

Low Throughput
High Latency
High Cost

Small models for Social-bias detector





"I will not hire someone who is an orange person."

unaligned AI assistant

- social bias, but not HAP

"Saying that you will not hire someone because they are orange is a harmful and stereotypical notion."

aligned AI assistant

- not social bias and not HAP

How well can small models perform?

model	FPR	FNR	Avg Err	# params
`gpt-3.5-instruct-turbo`*	25.56	13.33	19.44	?
`gpt-3.5-turbo` (ChatGPT 3.5)*	11.11	22.22	16.67	?
`gpt-4`*	8.89	20.00	14.44	?
`Llama-Guard`	12.22	20.00	16.11	7B
`Llama-Guard-2`	***4.44***	***26.67***	***15.56***	8B
`toxigen-hatebert`	53.33	12.22	32.78	0.1B
`social-bias`	90.00	18.89	54.44	0.1B

High false positives

Table 5: Results on Evaluation Set with 180 examples, * denotes results taken from [14]

NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps

Kristina Gligorić **Myra Cheng** **Lucia Zheng** **Esin Durmus** **Dan Jurafsky**
Stanford University
gligoric@cs.stanford.edu

How can we improve small models? – *Synthetic Data*

You are asked to come up with a set of {num_samples} diverse examples for the topic of {topic}, specifically focusing on {sub_topic}.

Please follow these guiding principles when generating responses:

1. Use proper grammar and punctuation.
2. The examples should be diverse in both style and form.
3. Each of the examples should be in a JSONL format.
4. Each example contains a biased statement and an unbiased statement.

To better assist you with this task, here is an example of the JSONL formatting:
{{
"biased": harmful and biased statement with a stereotype relating to {sub_topic},
"unbiased": unbiased and benign statement with no stereotypes relating to {sub_topic}
}}

Generate {num_samples} such examples for the topic of {topic}, specifically focusing on {sub_topic}. Remember to follow the principles mentioned above.

Return your responses below in JSONL format.

biased ⟶ use	unbiased ⟶ mention
All [redacted] people are fortune tellers and rely on trickery to make a living.	Some individuals from the [redacted] community may practice fortune telling, while others have different occupations and interests.
You can't trust someone with [redacted] nia because they hear voices and can't tell what's real.	People with [redacted] may experience auditory hallucinations, but they can still be trustworthy and reliable individuals with the right treatment and support.
[redacted] are superficial and materialistic.	[redacted] have different values and priorities, and some may emphasize spiritual growth, social justice, and environmental sustainability.

Manish Nagireddy, Inkit Padhi, Soumya Ghosh, Prasanna Sattigeri. *When in Doubt, Cascade: Towards Building Efficient and Capable Guardrails (Work in Progress)*

How can we improve small models? – *Synthetic Data*

model	FPR	FNR	Avg Err	# params
`gpt-3.5-instruct-turbo`*	25.56	13.33	19.44	?
`gpt-3.5-turbo` (ChatGPT 3.5)*	11.11	22.22	16.67	?
`gpt-4`*	8.89	20.00	14.44	?
`Llama-Guard`	12.22	20.00	16.11	7B
`Llama-Guard-2`	***4.44***	***26.67***	***15.56***	8B
`toxigen-hatebert`	53.33	12.22	32.78	0.1B
`social-bias`	90.00	18.89	54.44	0.1B
`social-bias-use-mention`	34.44	23.33	28.89	0.1B
`social-bias-use-mention-onetrial`	32.22	37.78	35.00	0.1B
`social-bias-onetrial-concat`	72.22	20.00	46.11	0.1B
`cascade-orig`	32.22	33.33	32.78	0.1B x 2
`cascade-onetrial`	26.67	47.78	37.22	0.1B x 2
`social-bias-distil`	95.56	5.56	50.56	0.04B
`social-bias-use-mention-distil`	26.67	24.44	25.56	0.04B
`cascade-distil`	**24.44**	**28.89**	**26.67**	0.04B x 2

Table 5: Results on Evaluation Set with 180 examples, * denotes results taken from [14]

false negative = missed flag!

much fewer parameters + competitive performance!

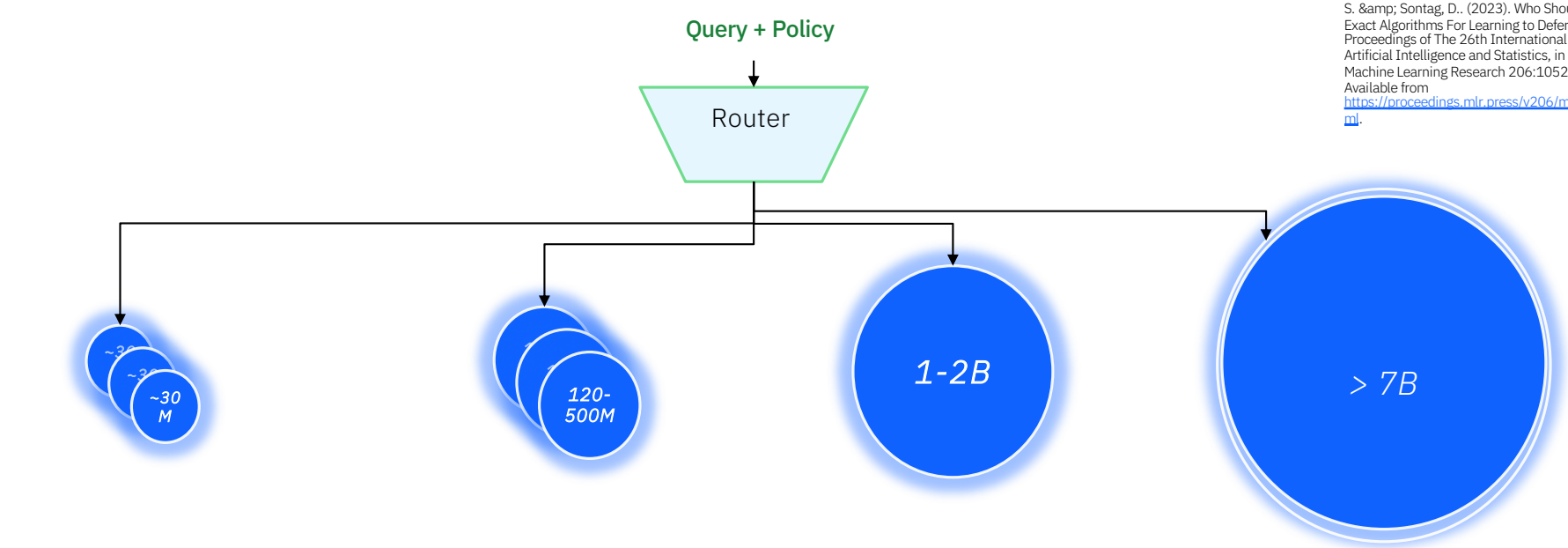
NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps

Kristina Gligorić **Myra Cheng** **Lucia Zheng** **Esin Durmus** **Dan Jurafsky**
Stanford University
gligoric@cs.stanford.edu

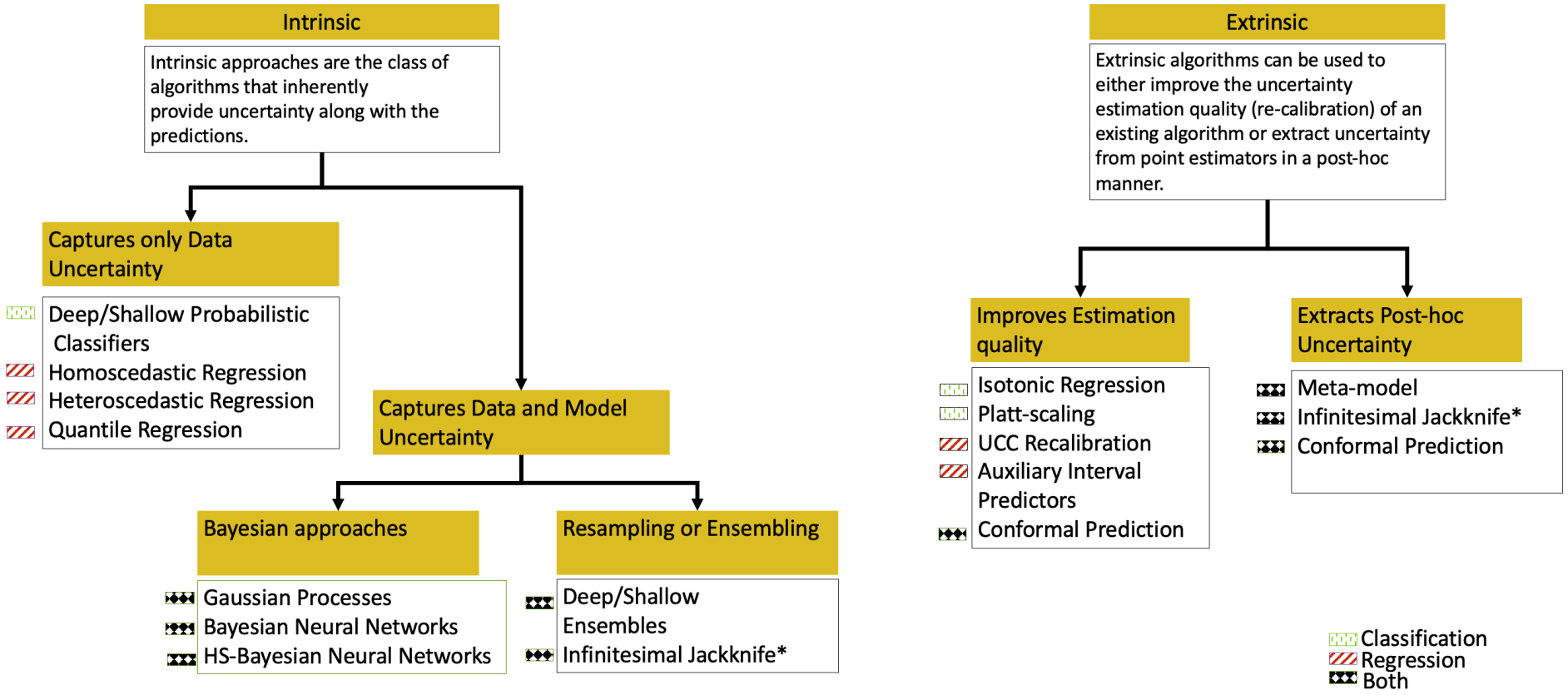
Policy Orchestration

Orchestrate based on cost/latency/accuracy/throughput requirements between different *sizes of models*.
Can be through *user defined policies or automated policies* [1]

[1] Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S. & Sontag, D.. (2023). Who Should Predict? Exact Algorithms For Learning to Defer to Humans. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 206:10520-10545 Available from https://proceedings.mlr.press/v206/mozannar23a.html.

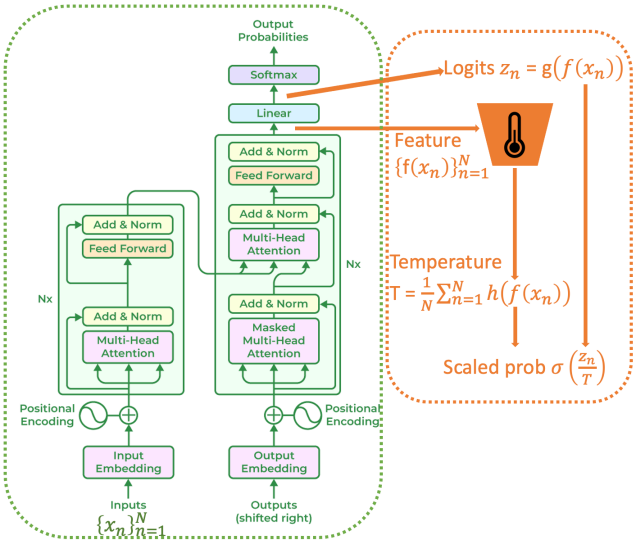


Ways to get uncertainty scores



Thermometer – amortized temperature scaling

$$f_\theta : R^D \rightarrow R_+$$



- f_θ is trained on related datasets and applied to a given dataset
- Light weight MLP, need one per LLM (frozen), needs to be trained once
- Works with decoder only / encoder-decoder / encoder only models

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. "Thermometer: Towards Universal Calibration for Large Language Models." In *Forty-first International Conference on Machine Learning*.

Training Datasets for Thermometer

MCQA

1.AG News dataset (converted to qa)
2.ARC-Easy
3.CODAH
4.Commonsense QA
5.COPA
6.Cosmos QA
7.DREAM
8.Fig QA
9.MedMCQA
10.Openbook QA
11.PIQA
12.RiddleSense
13.Social IQa

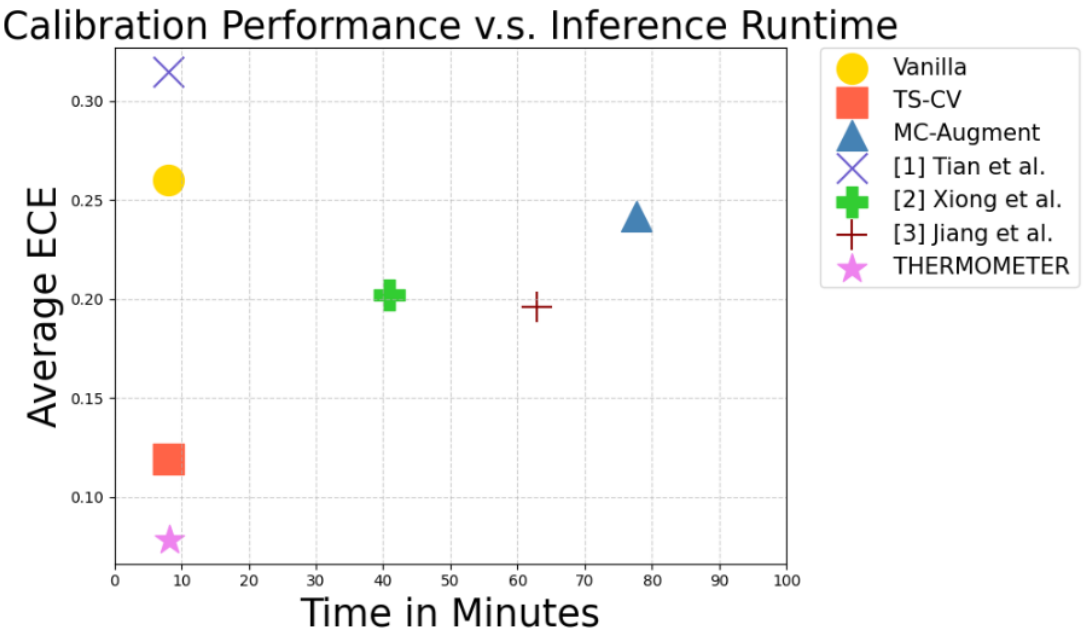
Big Bench

1.Winowhy
2.Strategyqa
3.play_dialog_same_or_different
4.movie_dialog_same_or_different
5.logical_fallacy_detection
6. contextual_parametric_knowledge_conflicts
7. epistemic_reasoning
8. fact_checker
9. formal_fallacies_syllogisms_negation
10. hyperbaton
11. bbq_lite_json
12. vitaminc_fact_verification
13. goal_step_wikihow
14. elementary_math_qa
15. unit_conversion
16. tracking_shuffled_objects

Free text QA

1.SQuAD.
2.NewsQA
3.TriviaQA
4.SearchQA
5.HotpotQA
6.NaturalQuestions
7.SciQA
8.Smiles QA
9.Web Questions
10.BioASQ
11.Race
12.DuoRC
13.DropQA
14.TextbookQA

Thermometer Results



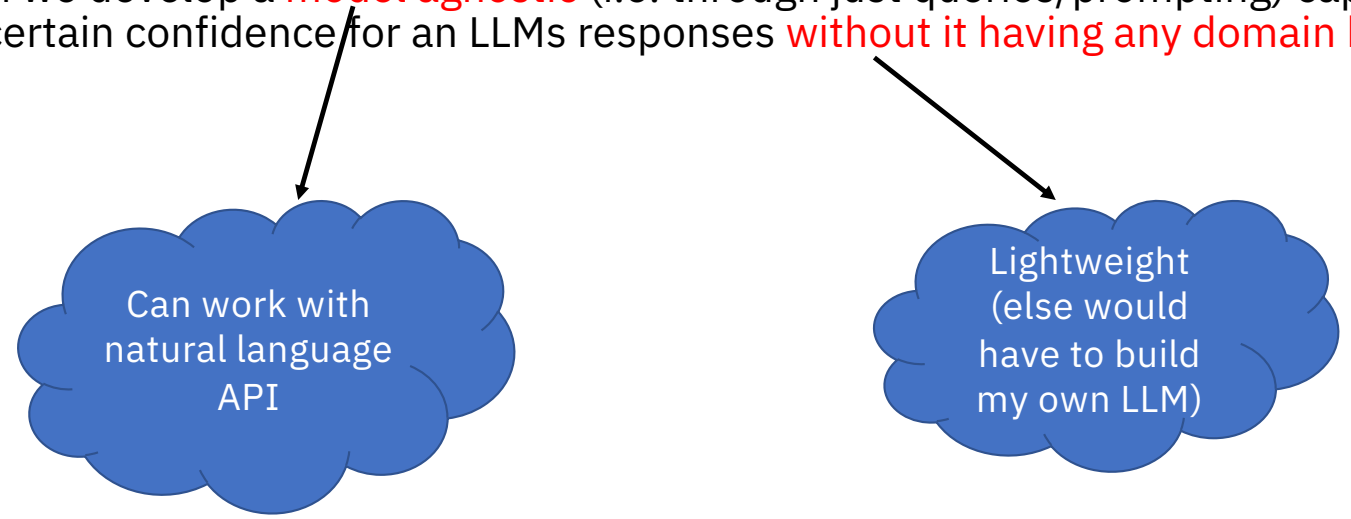
Above results with LLama-7b-chat on MMLU. Similar trends for other models/datasets

Tian et al., EMNLP 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback.
Xiong et al., ICLR 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs
Jiang et al., ICML 2023 Workshop. Calibrating Language Models via Augmented Prompt Ensembles.

Meta-model Framework

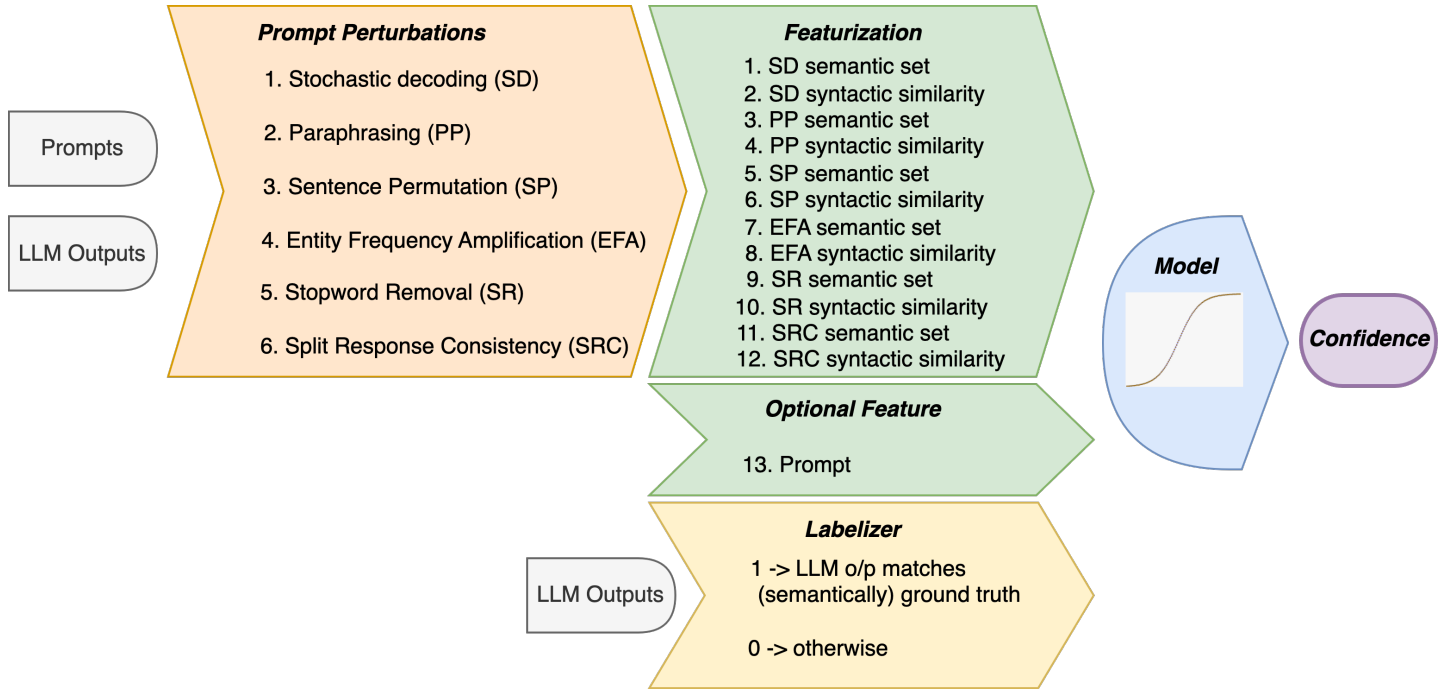
Question

Can we develop a model agnostic (i.e. through just queries/prompting) capability that can ascertain confidence for an LLMs responses without it having any domain knowledge?



Tejaswini Pedapati, Amit Dhurandhar, Soumya Ghosh, Soham Dan, and Prasanna Sattigeri. "Large Language Model Confidence Estimation via Black-Box Access." *arXiv e-prints* (2024): arXiv-2406.

Meta-model Framework



Goal: to integrate various prompt-perturbation methods and learn a calibrated combined score.

Prompt Perturbation Examples

- SD – does not change prompt
- PP – paraphrases
- SP – permutes order of sentence
- EFA – repeats entities
- SR- removes stop words
- SRC – does not change prompt

Prompt Perturbations

1. Stochastic decoding (SD)
2. Paraphrasing (PP)
3. Sentence Permutation (SP)
4. Entity Frequency Amplification (EFA)
5. Stopword Removal (SR)
6. Split Response Consistency (SRC)

Input Prompt
context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. They descended from the Normands ("Norman" comes from "Norseman") of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. During generations of assimilation and mixing with the native French and Roman-Gaulese populations, their descendants would gradually merge with the Carolingian cultures of West France. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed.
question: In what country is Normandy located?

Prompt Pert.	Perturbed Prompt	Output
SD		France, Denmark, Iceland, Norway
PP	context: Normandy, a region in France came to bear because of Normans in the 10th and 11th centuries. They descended from the Normands ~~("Norman" comes from "Norseman")~~ of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. There was generations of mixing with the Roman-Gaulese populations and native French. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed. question: In what country is Normandy located?	Iceland
SP	context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed. They descended from the Normands ("Norman" comes from "Norseman") of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. During generations of assimilation and mixing with the native French and Roman-Gaulese populations, their descendants would gradually merge with the Carolingian cultures of West France. question: In what country is Normandy located?	Denmark
EFA	context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. They descended from the Normands ("Norman" comes from "Norseman") of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. During generations of assimilation and mixing with the native French and Roman-Gaulese populations, their descendants would gradually merge with the Carolingian cultures of West France. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed. question: In what country is Normandy located?	France
SR	context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) ~~are~~ the people ~~who, in the~~ 10th and 11th centuries, gave their name ~~to~~ Normandy, a region of France. They descended ~~from the~~ Normands ("Norman" comes from "Norseman") ~~of the~~ raiders and pirates of Denmark, Iceland and Norway ~~who~~, under their leader Rollo, agreed ~~to~~ swear allegiance ~~to~~ King Charles III ~~of~~ France of the West. During generations ~~of~~ assimilation and mixing ~~with the~~ native French and Roman-Gaulese populations, ~~their~~ descendants would gradually merge ~~with the~~ Carolingian cultures ~~of~~ West France. ~~The~~ distinct cultural and ethnic identity ~~of the~~ Normans originally emerged ~~in the~~ first half ~~of the~~ 10th century, and it continued ~~to~~ evolve ~~over the~~ centuries that followed. question: In what country is Normandy located?	Norway
SRC		Normandy is located in Denmark. Normandy is located in Iceland.

Table 1: Below we see examples of different prompt perturbations for a prompt from the SQuAD dataset. The color blue and strike outs indicate changes to the input prompt. i) SD does not change the prompt (hence empty cell), but using a stochastic decoding scheme samples multiple responses (four example samplings shown). PP paraphrases the prompt. SP randomly reorders some of the sentences. EFA repeats certain sentences with entities in them. SR removes stopwords. SRC checks for consistency in reasonable size random splits of the LLM response (again prompt is not perturbed). The splitting of the two sentences indicates inconsistency as depicted in red. Thus, the perturbations test an LLM in complimentary ways.

Input Prompt
context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. They descended from the Normands ("Norman" comes from "Norseman") of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. During generations of assimilation and mixing with the native French and Roman-Gaulese populations, their descendants would gradually merge with the Carolingian cultures of West France. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed.
question: In what country is Normandy located?

Prompt Pert.	Perturbed Prompt	Output
SD		France, Denmark, Iceland, Norway
PP	context: Normandy, a region in France came to bear because of Normans in the 10th and 11th centuries. They descended from the Normands ~~("Norman" comes from "Norseman")~~ of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. There was generations of mixing with the Roman-Gaulese populations and native French. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed. question: In what country is Normandy located?	Iceland
SP	context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed. They descended from the Normands ("Norman" comes from "Norseman") of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. During generations of assimilation and mixing with the native French and Roman-Gaulese populations, their descendants would gradually merge with the Carolingian cultures of West France. question: In what country is Normandy located?	Denmark
EFA	context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) are the people who, in the 10th and 11th centuries, gave their name to Normandy, a region of France. They descended from the Normands ("Norman" comes from "Norseman") of the raiders and pirates of Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear allegiance to King Charles III of France of the West. During generations of assimilation and mixing with the native French and Roman-Gaulese populations, their descendants would gradually merge with the Carolingian cultures of West France. The distinct cultural and ethnic identity of the Normans originally emerged in the first half of the 10th century, and it continued to evolve over the centuries that followed. question: In what country is Normandy located?	France
SR	context: The Normans (Norman : Nourmands ; French : Normands ; Latin : Normanni) ~~are~~ the people ~~who, in the~~ 10th and 11th centuries, gave their name ~~to~~ Normandy, a region of France. They descended ~~from the~~ Normands ("Norman" comes from "Norseman") ~~of the~~ raiders and pirates of Denmark, Iceland and Norway ~~who~~, under their leader Rollo, agreed ~~to~~ swear allegiance ~~to~~ King Charles III ~~of~~ France of the West. During generations ~~of~~ assimilation and mixing ~~with the~~ native French and Roman-Gaulese populations, ~~their~~ descendants would gradually merge ~~with the~~ Carolingian cultures ~~of~~ West France. ~~The~~ distinct cultural and ethnic identity ~~of the~~ Normans originally emerged ~~in the~~ first half ~~of the~~ 10th century, and it continued ~~to~~ evolve ~~over the~~ centuries that followed. question: In what country is Normandy located?	Norway
SRC		Normandy is located in Denmark. Normandy is located in Iceland.

Featurization

Featurization

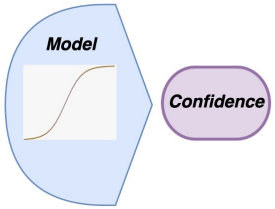
1. SD semantic set
2. SD syntactic similarity
3. PP semantic set
4. PP syntactic similarity
5. SP semantic set
6. SP syntactic similarity
7. EFA semantic set
8. EFA syntactic similarity
9. SR semantic set
10. SR syntactic similarity
11. SRC semantic set
12. SRC syntactic similarity

semantic set feature: number of semantically similar outputs based on bi-directional entailment; lower number indicating more confidence.
Syntactic/lexical similarity: Rouge, BLEU, etc
SRC minimum: lowest entailment score across different parts of the each response

Experiments (Confidence Estimation)

Dataset(LLM)	# of Semantic Sets	Lexical Similarity	EigenValue	Eccentricity	Degree	Ours
TriviaQA(llama)	0.73	0.76	0.77	0.76	0.77	**0.88**
TriviaQA(flan-ul2)	0.83	0.8	0.86	0.86	0.87	**0.95**
TriviaQA(mistral)	0.65	0.72	0.76	0.75	0.75	**0.81**
SQuAD(llama)	0.65	0.72	0.74	0.58	0.72	**0.83**
SQuAD(flan-ul2)	0.6	0.7	0.67	0.65	0.67	**0.8**
SQuAD(mistral)	0.59	0.7	0.67	0.65	0.67	**0.84**
CoQA(llama)	0.61	0.74	0.76	0.76	0.77	**0.92**
CoQA(flan-ul2)	0.61	0.76	0.78	0.78	0.79	**0.87**
CoQA(mistral)	0.56	0.74	0.79	0.77	0.79	**0.81**
NQ(llama)	0.65	0.75	0.75	0.73	0.74	**0.85**
NQ(flan-ul2)	0.76	0.76	0.86	0.86	0.86	**0.93**
NQ(mistral)	0.66	0.73	0.77	0.77	0.78	**0.83**

AUROC Results



Train confidence model based on the described features where labels are generated depending on if the LLM output matches the ground truth based on rouge (thresholded).

LLMs: flanul2, llama-13b and mistral-7b

Datasets: TriviaQA, SquAD, CoQA and NQ

Train Size: 1000 examples (same dataset size used for tuning in previous works).

Dataset(LLM)	# of Semantic Sets	Lexical Similarity	EigenValue	Eccentricity	Degree	Ours
TriviaQA(llama)	0.77	0.8	0.8	0.8	0.8	**0.83**
TriviaQA(flan-ul2)	0.69	0.72	0.73	0.73	0.73	**0.74**
TriviaQA(mistral)	0.55	0.63	**0.64**	**0.64**	**0.64**	**0.64**
SQuAD(llama)	0.3	0.36	0.37	0.28	0.36	**0.68**
SQuAD(flan-ul2)	0.73	0.95	0.83	0.82	0.83	**0.96**
SQuAD(mistral)	0.72	0.93	0.82	0.82	0.82	**0.96**
CoQA(llama)	0.56	0.67	0.67	0.67	0.67	**0.71**
CoQA(flan-ul2)	0.7	0.79	**0.8**	0.79	0.79	**0.8**
CoQA(mistral)	0.46	0.62	0.64	0.63	**0.64**	0.49
NQ(llama)	0.37	0.41	0.42	0.41	0.41	**0.45**
NQ(flan-ul2)	0.41	0.44	**0.47**	0.46	0.45	**0.47**
NQ(mistral)	0.32	0.38	0.40	0.40	0.39	**0.42**

AUARC Results

Experiments (Interpretability)

Dataset(LLM)	Rank 1	Rank 2	Rank 3	Rank 4
TriviaQA(llama)	SD syntactic similarity	SD semantic set	SR syntactic similarity	PP syntactic similarity
TriviaQA(flan-ul2)	SD syntactic similarity	SD semantic set	PP semantic set	PP syntactic similarity
TriviaQA(mistral)	SD syntactic similarity	PP syntactic similarity	SP semantic set	SD semantic set
SQuAD(llama)	SP syntactic similarity	EFA semantic set	-	-
SQuAD(flan-ul2)	SP syntactic similarity	-	-	-
SQuAD(mistral)	SP syntactic similarity	EFA semantic set	-	-
CoQA(llama)	SD syntactic similarity	EFA semantic set	SD semantic set	SR syntactic similarity
CoQA(flan-ul2)	SD syntactic similarity	EFA semantic set	SD semantic set	SP syntactic similarity
CoQA(mistral)	SD syntactic similarity	SD semantic set	EFA semantic set	EFA syntactic similarity
NQ(llama)	PP syntactic similarity	SD semantic set	SD syntactic similarity	SP syntactic similarity
NQ(flan-ul2)	SR semantic set	SD syntactic similarity	SP syntactic similarity	PP syntactic similarity
NQ(mistral)	PP syntactic similarity	SD semantic set	SD syntactic similarity	SP syntactic similarity

Prompt Perturbations

1. Stochastic decoding (SD)
2. Paraphrasing (PP)
3. Sentence Permutation (SP)
4. Entity Frequency Amplification (EFA)
5. Stopword Removal (SR)
6. Split Response Consistency (SRC)

Top features consistent across models but vary across datasets.

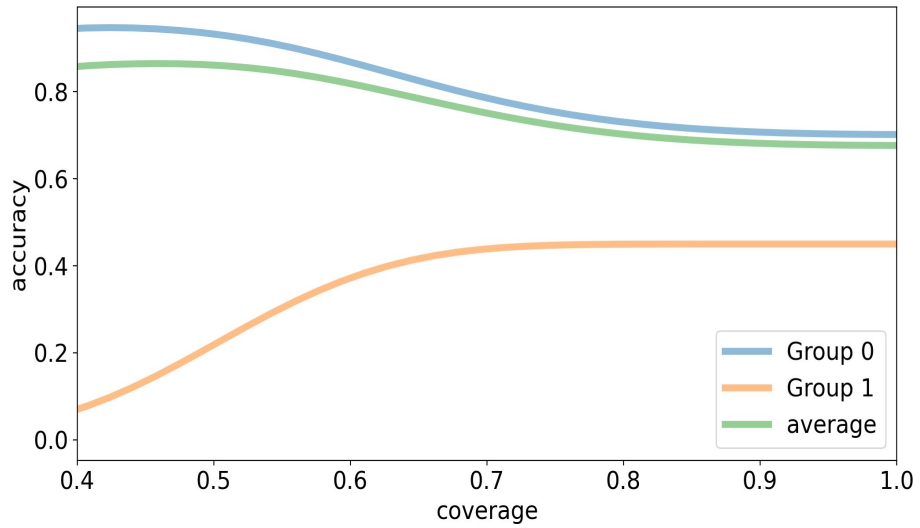
Experiments (Transferability)

Dataset	Source LLM	AUROC Self	Target LLM 1 AUROC	Target LLM 2 AUROC
TriviaQA	llama	0.88	0.94 (flan-ul2)	0.80 (mistral)
	flan-ul2	0.94	0.87 (llama)	0.80 (mistral)
	mistral	0.81	0.84 (llama)	0.91 (flan-ul2)
SQuAD	llama	0.83	0.81 (flan-ul2)	0.80 (mistral)
	flan-ul2	0.8	0.79 (llama)	0.78 (mistral)
	mistral	0.84	0.82 (llama)	0.83 (flan-ul2)
CoQA	llama	0.92	0.79 (flan-ul2)	0.78 (mistral)
	flan-ul2	0.87	0.87 (llama)	0.81 (mistral)
	mistral	0.81	0.88 (llama)	0.86 (flan-ul2)
NQ	llama	0.85	0.91 (flan-ul2)	0.83 (mistral)
	flan-ul2	0.93	0.83 (llama)	0.82 (mistral)
	mistral	0.83	0.85 (llama)	0.90 (flan-ul2)

Dataset	Source LLM	AUARC Self	Target LLM 1 AUARC	Target LLM 2 AUARC
TriviaQA	llama	0.83	0.74 (flan-ul2)	0.64 (mistral)
	flan-ul2	0.74	0.83 (llama)	0.64 (mistral)
	mistral	0.64	0.83 (llama)	0.73 (flan-ul2)
SQuAD	llama	0.68	0.62 (flan-ul2)	0.63 (mistral)
	flan-ul2	0.96	0.89 (llama)	0.91 (mistral)
	mistral	0.96	0.90 (llama)	0.91 (flan-ul2)
CoQA	llama	0.71	0.79 (flan-ul2)	0.49 (mistral)
	flan-ul2	0.80	0.70 (llama)	0.49 (mistral)
	mistral	0.49	0.69 (llama)	0.79 (flan-ul2)
NQ	llama	0.45	0.46 (flan-ul2)	0.42 (mistral)
	flan-ul2	0.47	0.45 (llama)	0.42 (mistral)
	mistral	0.42	0.45 (llama)	0.46 (flan-ul2)

Bias in Selective Prediction

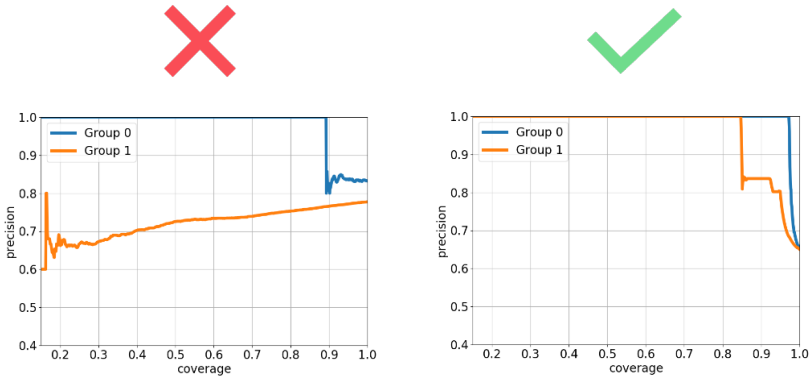
Even though we may have good average selective prediction performance but perform poorly on certain groups, where reducing uncertainty threshold may result in a decrease in performance for under-represented group.



Fair measure of Uncertainty Quality

monotonic selective risk (MSR)

MSR requires our predictor and uncertainty measure to ensure that the subgroup error rate (or MSE for regression) decreases *monotonically* with a decrease in coverage for *every* subgroup.



Another possible fairness notion is to require same performance and coverage curve for all subgroups.

But it is still possible that the error rate of both subgroups increases when coverage decreases, which violates the primary goal of selective prediction!

Joshua K. Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W. Wornell. "Fair selective classification via sufficiency." In *International conference on machine learning*, pp. 6076-6086. PMLR, 2021.

Generic risks vs. specific risks

Common across sectors and use cases

- poor predictive accuracy
- lack of fairness and equity
- lack of explainability
- model uncertainty
- distribution shifts (drift)

- poisoning attacks
- evasion attacks
- extraction attacks
- inference attacks
- model transparency

- inability to reason
- privacy leakage
- prompt injection attacks
- misinformation

- hallucinations
- lack of factuality or faithfulness
- lack of source attribution
- toxicity, profanities, and hate speech
- bullying and gaslighting

Unique or particular to a company

laws

industry standards

social norms of end-users

corporate policies

market demands

technology architecture constraints

Major Financial Institution: "*Do not include the name of any competitor bank in chatbot responses."*

Major Chicago Restaurant: "*Conversations must be snarky!"*

IBM Business Conduct Guidelines: "*Never offer or give anyone, or accept from anyone, anything of value that is, or could be viewed as, a bribe, kickback or other improper benefit, and never improperly attempt to influence that person's or entity's relationship with IBM, whether to obtain or retain business or get some other benefit."*

Thank you!