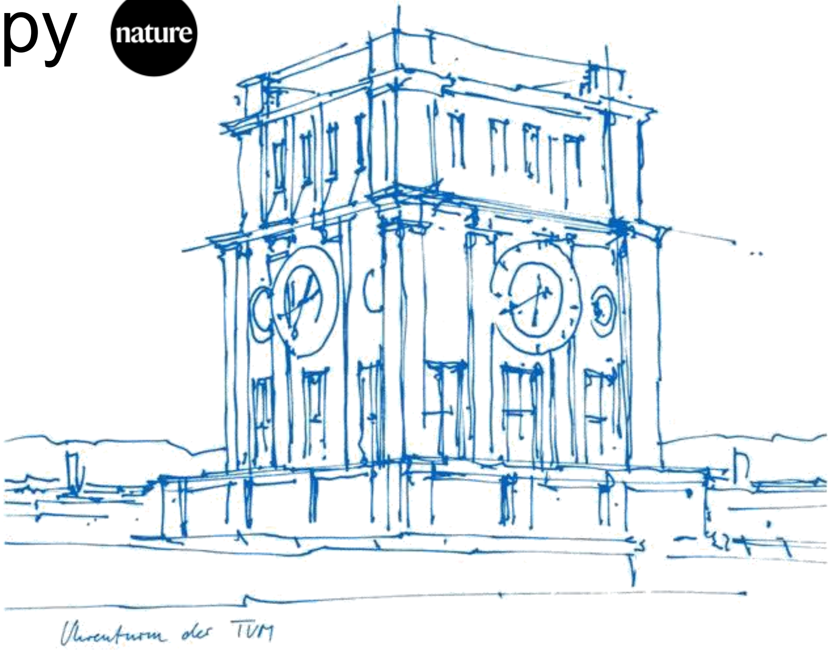
Chair of Computational Imaging and AI in Medicine
TUM School of Computation, Information and Technology
Technical University of Munich

Detecting hallucinations in large language models using semantic entropy

by Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn & Yarin Gal (June 2024)



Francesco Vaccaro

Garching near Munich, 2 July 2025

Outline

- Motivation: Hallucinations & Confabulations
- Theoretical foundation: Naive entropy vs. semantic entropy (SE)
- SE with different output lengths
- Performance comparison
- Discussion: Strengths & Limitations

Motivation: Hallucinations

[Pubity]

Altman on the OpenAI Podcast (June 18, 2025):

"People have a **very high degree of trust** in ChatGPT, which is interesting because […] **AI hallucinates**"

"[AI] should be the tech that you **don't trust** that much"

Hallucinations & Confabulations

[Smith+, 2023]

	Clinical	LLMs
Hallucinations	• sensory experiences without respective external stimuli	• umbrella term for wrong outputs
Confabulations	• generation of narrative details • details are incorrect & not recognized as such	• subtype of Hallucinations: wrong and **arbitrary**

Confabulations vs. other Hallucinations

Causes of other hallucination types	**Example**
erroneous training data	common misconceptions like “Napoleon was small”
LLM lies in pursuit of a reward	“generate titles for YT videos” & “optimize for clicks via RL”
systematic failure of reasoning or generalization	*Training*: Q: 3 Apples + 5 Apples A: 8 Apples ✓ *Inference*: Q: four Apples plus two Apples A: four Apples ✗

[Evans+ 2021]

Naive vs. semantic entropy // short LLM outputs

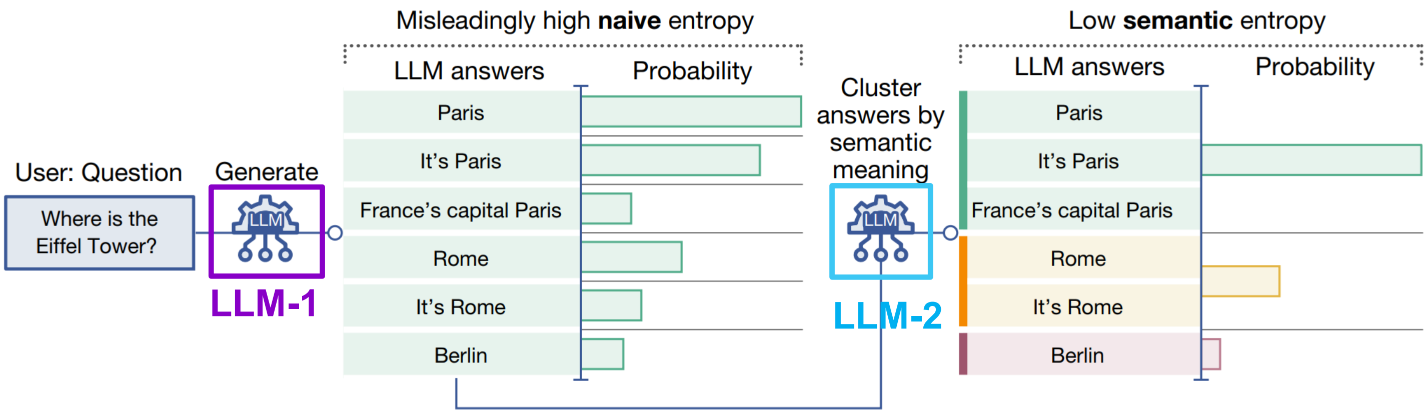
a) many different outputs

b) similar probabilities

→ **higher entropy**

	naive entropy	semantic entropy
Basis for entropy calculation	output token sequence	output meaning
How to determine	directly based on sequence probabilities	More complex: LLM-2 based clustering by meaning

Naive vs. semantic entropy // short LLM outputs



	naive entropy	semantic entropy
Basis for entropy calculation	output token sequence	output meaning
How to determine	directly based on sequence probabilities	More complex: LLM-2 based clustering by meaning

NE vs SE // short outputs // Select Examples

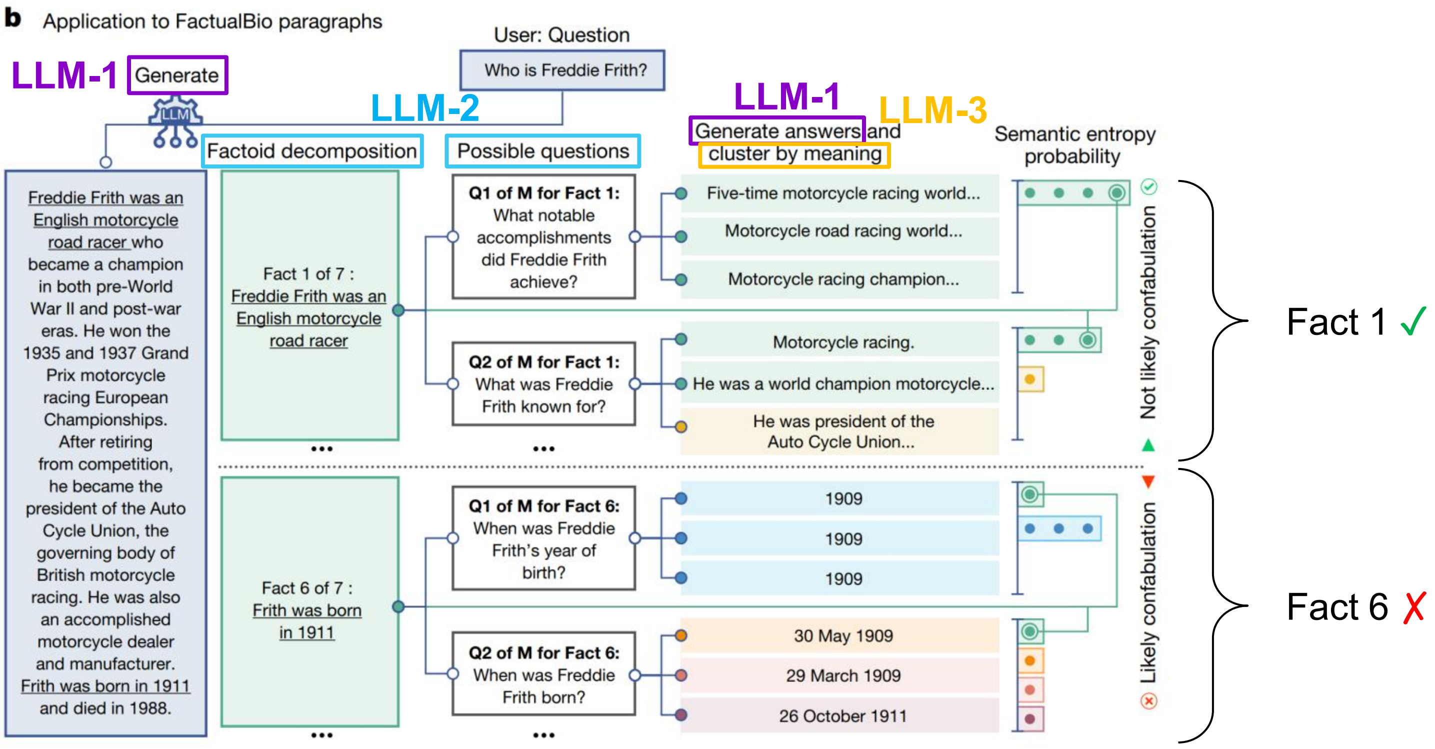
Question	Model answer	Reference answer	Generations for entropy	Prediction	
				Semantic entropy	Naive entropy
Refineries, process chemical, power generation, mills and manufacturing plants are under what sector of construction?	Refineries, process chemical, power generation, mills and manufacturing plants are under the industrial sector of construction.	Industrial or heavy civil or heavy engineering.	• All the above are under the industrial sector of construction. • The refineries, process chemical, power generation, mills and manufacturing plants are under the industrial sector of construction. • These are all under the heavy industrial sector of construction.	Not confabulation ✓	Confabulation ✗

Reference

"best answer"
(sampled at low Temperature T = 0.1)
→ assess model accuracy

- Same meaning (1 cluster)
- different token-sequences with similar probabilities

Semantic Entropy for longer LLM outputs

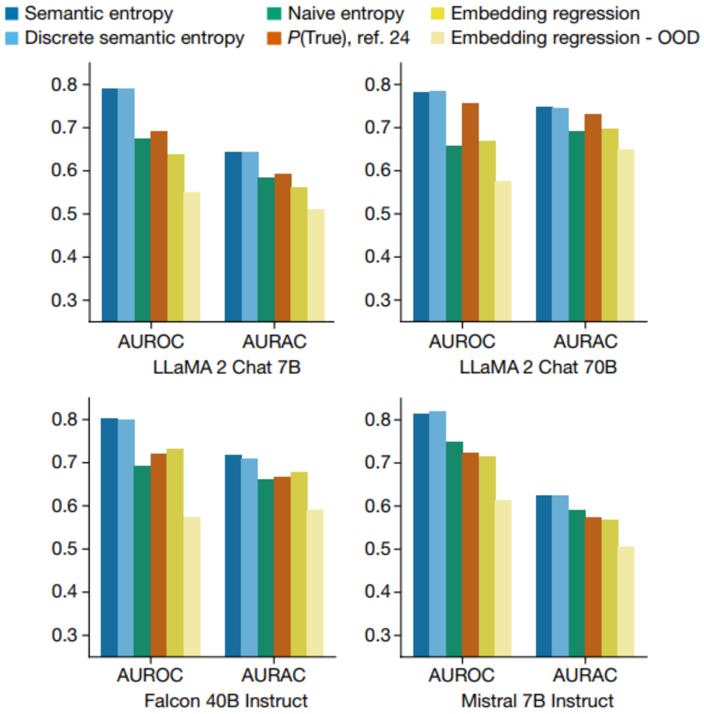


SE vs. other uncertainty metrics (short outputs)

- LLM answer correct?
  → treated as binary classification problem

- AUROC
  - measures quality of mistake prediction
  - ROC: FPR vs TPR

- AURAC
  - measures performance improvement with increasing rejection rate
  - basis: RAC = rejection accuracy curve
  - more sensitive to overall sensitivity of the model than AUROC



Discussion: Strengths of SE

- good performance
- "unsupervised" → no labels required
- no training → unsensitive to distribution shifts
- domain-independent (premise: working entailment estimator)
- discrete variant that works without model internals

Discussion: Limitations of SE

- SE misses confidently wrong answers (esp. non-confabulations)
- strong dependence on entailment model (NLI) accuracy
- Extra compute & time (sampling + NLI)
- not applicable for single generations during real usage (instead: RL?)

Takeaways

- SE is useful for flagging confabulations
- simple, probabilistic, training-free method → powerful uncertainty metric

"LLMs are [quite good] at **knowing what they don't know**
[, but they] **don't know they know what they don't know**."

[Kyle 2025]

Questions?

References

(1) [Farquhar+ 2024] (presented paper) (is always the source, when no other source is given)

(2) pubity [@pubity]. "On the first episode of OpenAI’s new podcast, CEO Sam Altman addressed something most people overlook, our growing trust in AI tools like ChatGPT. …" *Instagram*, 26 June 2025, www.instagram.com/p/DLVoy-rNiA3/. Accessed 29 June 2025.

(3) A. L. Smith, F. Greaves, and T. Panch, “Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models,” PLOS Digit Health, vol. 2, no. 11, p. e0000388, Nov. 2023, doi: 10.1371/journal.pdig.0000388.

(4) O. Evans et al., “Truthful AI: Developing and governing AI that does not lie,” Oct. 13, 2021, arXiv: arXiv:2110.06674. doi: 10.48550/arXiv.2110.06674.

(5) Barr, Kyle. "Sam Altman’s Lies About ChatGPT Are Growing Bolder." *Gizmodo*, 11 June 2025, www.gizmodo.com/sam-altmans-lies-about-chatgpt-are-growing-bolder-2000614431. Accessed 29 June 2025.

Additional Material

SE vs. other uncertainty metrics (short outputs)

Method	Labels required?	Training necessary?	# required generations	Details
Embedding regression [- OOD]	Yes	Yes	1	Logistic regression, trained with labels [trained on different distribution]
p(True) (few-shot)	Yes ("in-context")	No	1	Uses labeled examples in prompt
[discrete -] Semantic entropy	**No**	No	**Multiple**	No labels required (only for evaluation)

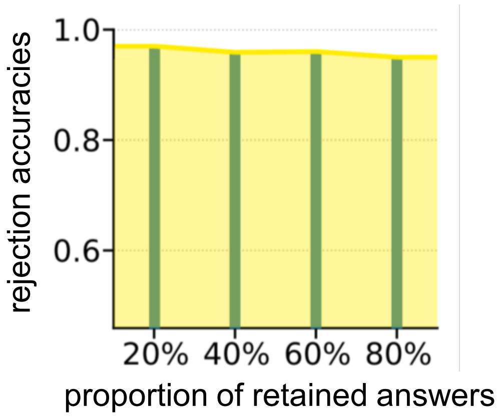
Performance metrics

- LLM answer correct?
  → treated as binary classification problem

- AUROC
  - measures quality of mistake prediction
  - ROC: FPR vs TPR

- AURAC
  - measures performance improvement with increasing rejection rate
  - basis: RAC = rejection accuracy curve
  - more sensitive to overall sensitivity of the model than AUROC

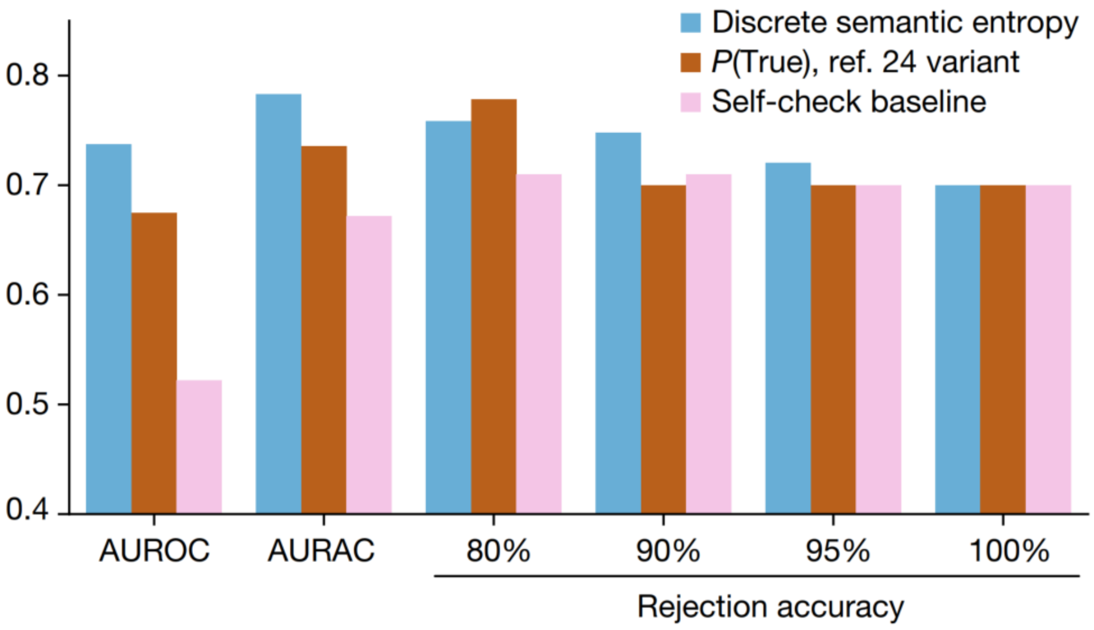
Exemplary RAC



Model: LLaMa 2 Chat 70 B
Uncertainty metric: Semantic Entropy
Dataset: TriviaQA

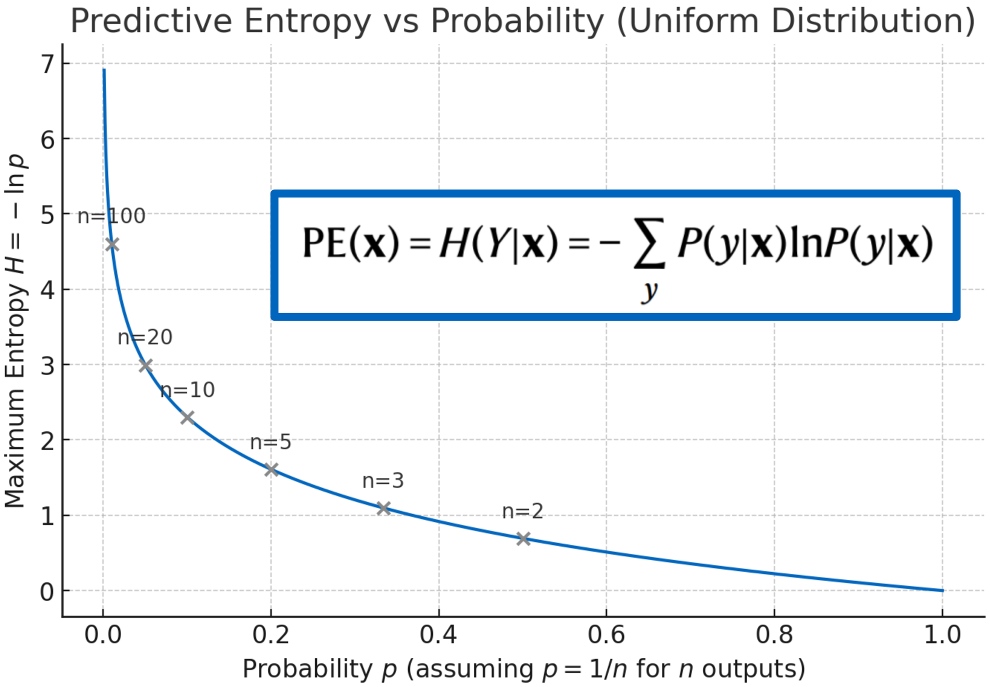
[Adapted after Supp. Fig. 6]

SE vs. other uncertainty metrics (long outputs)



(% of retained answers)

Maximum Predictive Entropy: Uniform Distribution



$$\boxed{\mathrm{PE}(\mathbf{x}) = H(Y|\mathbf{x}) = -\sum_{y} P(y|\mathbf{x}) \ln P(y|\mathbf{x})}$$