

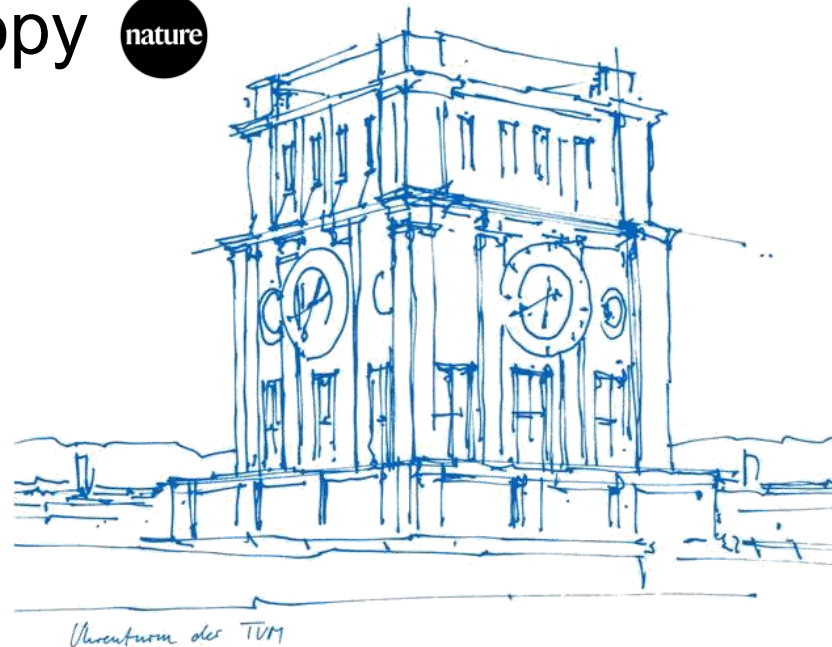
# Detecting hallucinations in large language models using semantic entropy

nature

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	<ul style="list-style-type: none"><li>• sensory experiences without respective external stimuli</li></ul>	<ul style="list-style-type: none"><li>• umbrella term for wrong outputs</li></ul>
Confabulations	<ul style="list-style-type: none"><li>• generation of narrative details</li><li>• details are incorrect &amp; not recognized as such</li></ul>	<ul style="list-style-type: none"><li>• subtype of Hallucinations: wrong and <b>arbitrary</b></li></ul>

# 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	<i>Training:</i> Q: 3 Apples + 5 Apples A: 8 Apples ✓ <i>Inference:</i> Q: four Apples plus two Apples A: four Apples ✗

# Naive vs. semantic entropy // short LLM outputs

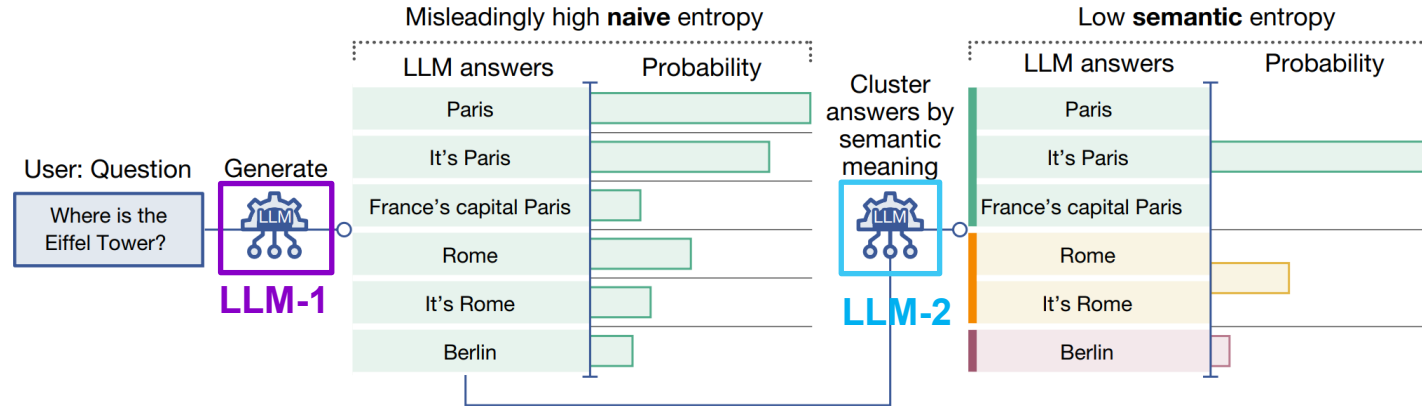
a) many different outputs  
b) similar probabilities



**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 <b>what sector of construction?</b>	Refineries, process chemical, power generation, mills and manufacturing plants are under the industrial sector of construction.	<b>Industrial</b> or heavy civil or heavy engineering.	<ul style="list-style-type: none"> <li>All the above are under the <b>industrial</b> sector of construction.</li> <li>The refineries, process chemical, power generation, mills and manufacturing plants are under the <b>industrial</b> sector of construction.</li> <li>These are all under the heavy <b>industrial</b> sector of construction.</li> </ul>	<u>Not confabulation</u> ✓ ↶ <i>Reference</i> ↷	<u>Confabulation</u> ✗

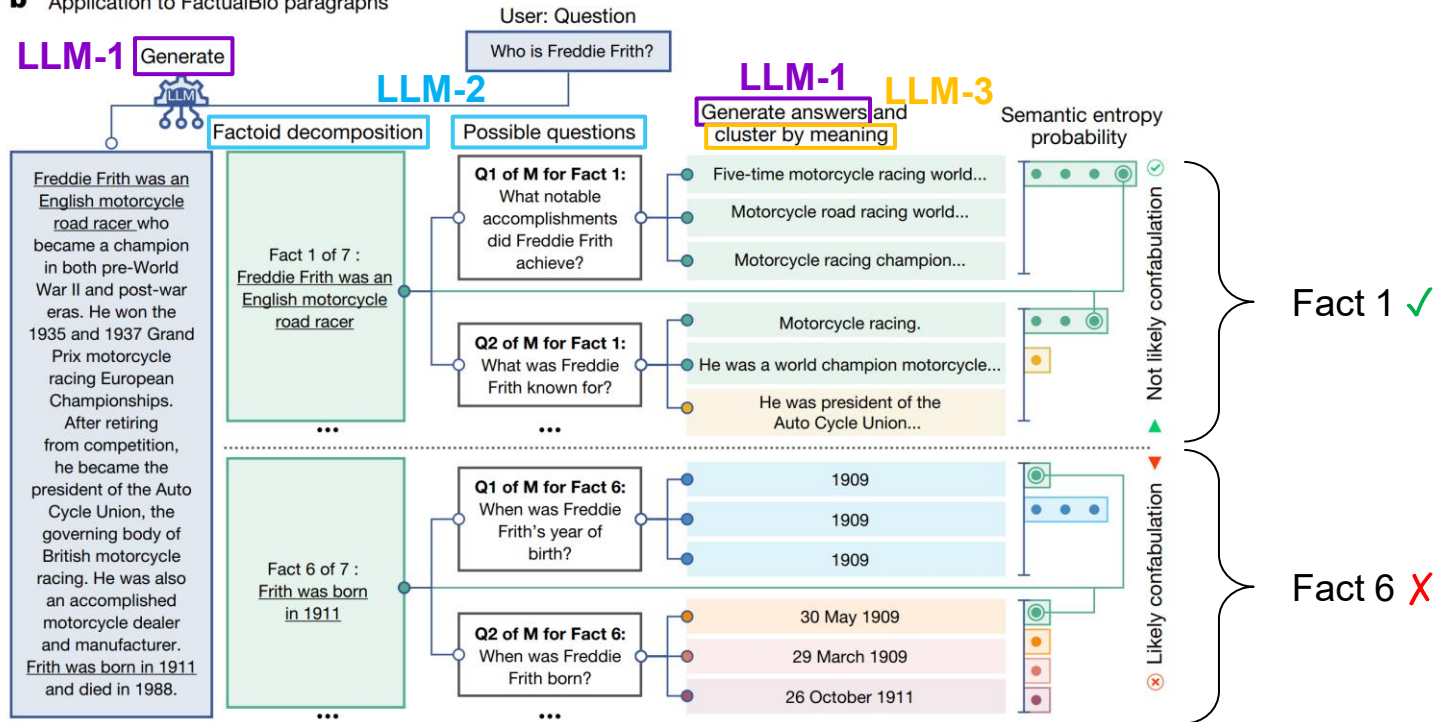
"best answer"  
(sampled at low Temperature  $T = 0.1$ )  
→ assess model accuracy

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



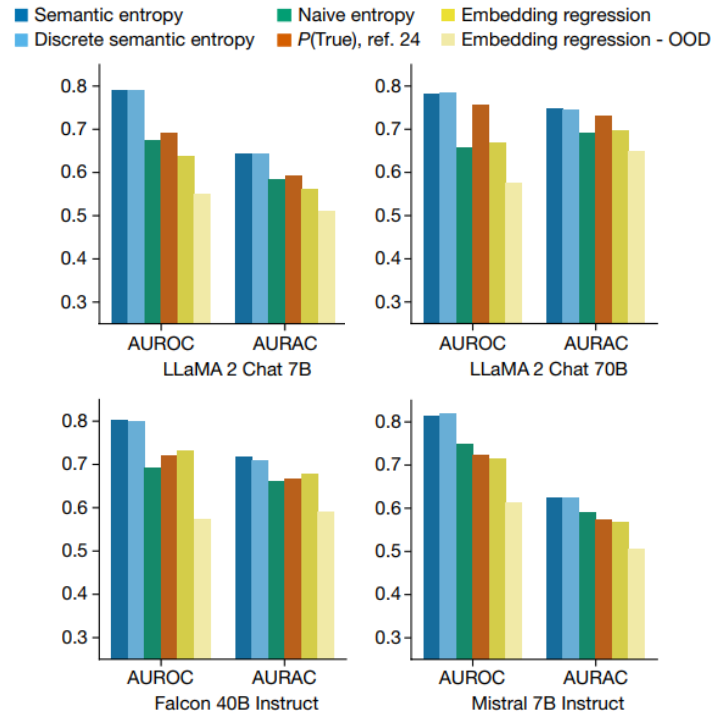
# Semantic Entropy for longer LLM outputs

## b Application to FactualBio paragraphs



# 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

Reference (1) is the presented papers and is always the source of figures, when no other source is given.

- (1) S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” *Nature*, vol. 630, no. 8017, pp. 625–630, Jun. 2024, doi: 10.1038/s41586-024-07421-0.
- (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/](https://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](https://www.gizmodo.com/sam-altmans-lies-about-chatgpt-are-growing-bolder-2000614431). Accessed 29 June 2025.

# Additional Material

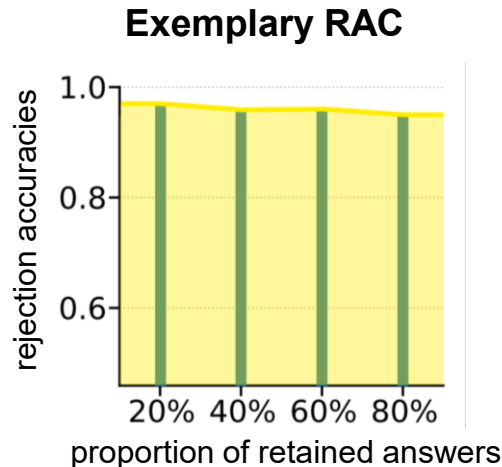


# SE vs. other uncertainty metrics (short outputs)

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

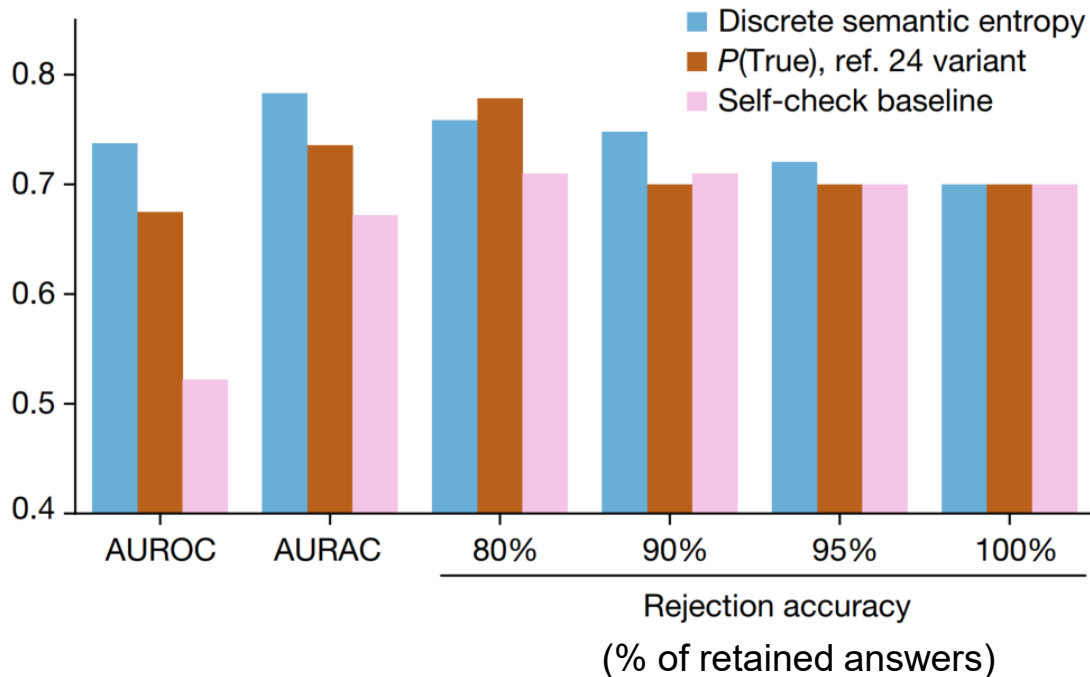


Model: LLaMa 2 Chat 70 B

Uncertainty metric: Semantic Entropy

Dataset: TriviaQA

# SE vs. other uncertainty metrics (long outputs)



# Maximum Predictive Entropy: Uniform Distribution

