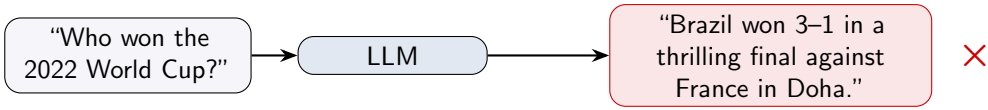
Hallucination & Grounding

Detection · Mitigation · Calibration · Attribution

What is hallucination?

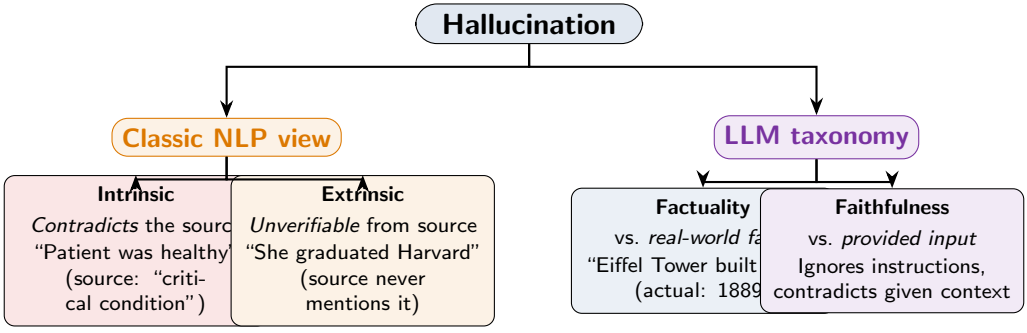


Hallucination: generated content that is **fluent, confident**, and **plausible** — but factually wrong or unsupported by any source.

Not random noise. Coherent, detailed fabrication.

The model doesn't "know" it's wrong — it produces the statistically most likely continuation.

Taxonomy of hallucinations



Maynez et al., ACL 2020

Huang et al., ACM TOIS 2024

Faithfulness sub-types: instruction inconsistency (wrong task) · context inconsistency (ignores RAG docs) · logical inconsistency (self-contradicts)

Types of hallucination

Factual errors

Confidently wrong facts

"The Great Wall is visible from space"

Fabricated citations

Non-existent references

GPT-3.5: 55% of generated citations are fake

Entity confusion

Merging attributes of different entities

Wrong person's discoveries

Temporal confusion

Wrong dates, mixing eras

Post-cutoff confabulation about recent events

Confabulation

Detailed fictional narratives presented as fact

Entirely invented backstories

The danger: hallucinations are not random — they are *coherent*, *detailed*, and *confident*, making them hard to detect without external verification.

Walters & Wilder, *Scientific Reports*, 2023 · Ji et al., ACM Computing Surveys, 2023

Why LLMs hallucinate — training & architecture

Training data

Web-scale corpora contain contradictions, falsehoods, and outdated information.

The model learns *all* of it.

Exposure bias

Training: sees ground-truth prefix
Inference: sees its *own* outputs

Small early error $\Rightarrow$ snowball

Objective mismatch

Trained to maximize
$P(\text{next token} \mid \text{context})$
not $P(\text{correct token})$

Fluency $\neq$ truthfulness

Softmax bottleneck

Hidden dim $d \ll \text{rank}(A)$
where A = true log-prob matrix

Yang et al., ICLR 2018

These are *structural* causes — they cannot be fully fixed by scaling alone.

Why LLMs hallucinate — generation dynamics

Temperature effects

$$P(x_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

High $T \to$ flat dist $\to$ random

$T = 0 \to$ greedy traps

Sweet spot: $T \in [0.3, 0.7]$

Knowledge cutoff

Training data has a fixed temporal boundary.

Post-cutoff questions $\Rightarrow$ plausible but fabricated answers

No world model

All "knowledge" is compressed into parameters — lossy.

No mechanism to distinguish *knowing* vs. *pattern-matching*

Knowledge conflicts

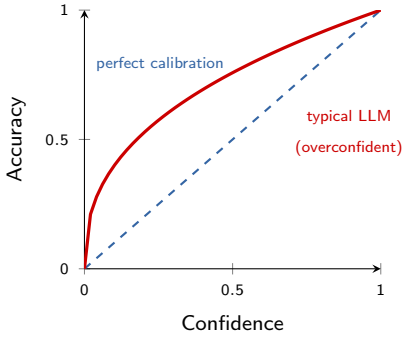
Parametric vs. contextual knowledge can disagree.

Popular entities: favor memory

Rare entities: favor context

Xie et al., EMNLP 2024

The confidence problem



$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|$$

Expected Calibration Error

RLHF makes it worse

Rewarding confident-sounding answers creates a confidence–accuracy gap

Verbalized confidence

"I'm 90% sure" is poorly calibrated — models default to high confidence regardless

Kadavath et al., 2022: "Language models (mostly) know what they know"

Real-world consequences

Legal

Mata v. Avianca (2023)

Lawyers used ChatGPT; 6 fabricated case citations.

Judge fined them $5,000.

Medical

Chatbots accepted false medical claims **32%** of the time.

Incorrect diagnoses risk delayed treatment.

Academic

Fabricated citations enter the literature.

"Hallucination laundering": LLM fictions become future training data.

Stanford study: LLMs hallucinate on legal queries **69–88%** of the time

GPT-3.5: 55% fabricated citations

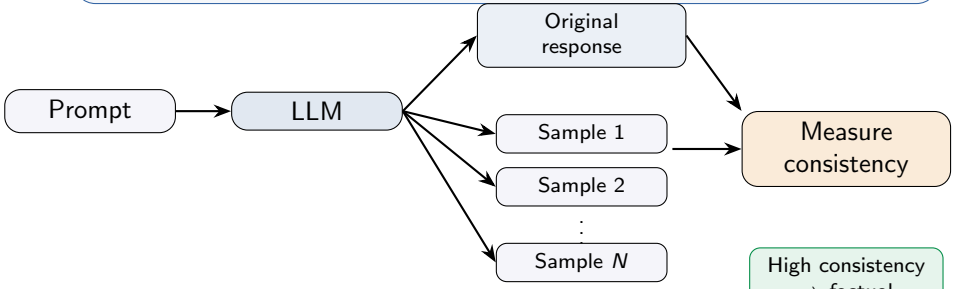
· **GPT-4:** 18% fabricated citations

Dahl et al., *Journal of Legal Analysis*, 2024

· Walters & Wilder, *Scientific Reports*, 2023

Detection: SelfCheckGPT

Key insight: if the model truly "knows" something, multiple stochastic samples will be **consistent**. Hallucinated facts will **vary** across samples.

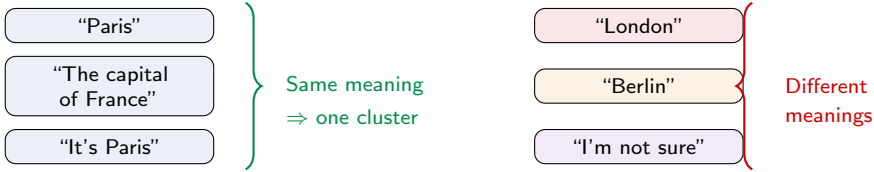


Zero-resource, **black-box** — no external databases, no model internals needed

Manakul et al., EMNLP 2023

Detection: Semantic Entropy

Problem: token-level entropy doesn't capture *meaning*.
"Paris" and "The capital of France" = different tokens, **same meaning**.



$$\text{SE} = -\sum_c P(c) \log P(c)$$

c = semantic clusters (not individual tokens)

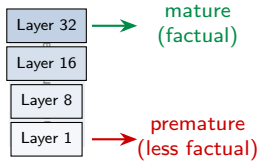
Low SE ⇒ consistent meaning ⇒ likely **correct**

Farquhar et al., Nature 2024

High SE ⇒ inconsistent meaning ⇒ likely **hallucinated**

Detection methods compared

DoLa: Decoding by Contrasting Layers



Contrast late
vs. early logits
$\Rightarrow$ amplify factual signal

Chuang et al., ICLR 2024

Method	Access	Key idea
SelfCheckGPT	Black-box	Multi-sample consistency
Semantic Entropy	Black-box	Meaning-level uncertainty
DoLa	White-box	Layer-contrastive decoding
Logit-based	White-box	Token probability thresholds
NLI-based	Black-box	Entailment checking vs. source

Mitigation: RAG as grounding

Without RAG

Prompt → LLM → Answer

Knowledge from parameters only.

May hallucinate freely.

With RAG

Prompt → Retrieve → LLM → Answer;

Grounded in retrieved docs.

Can cite sources.

Impact: GPT-4 + RAG achieved **0% hallucination** on cancer queries vs. 6% without RAG (JMIR Cancer, 2024)

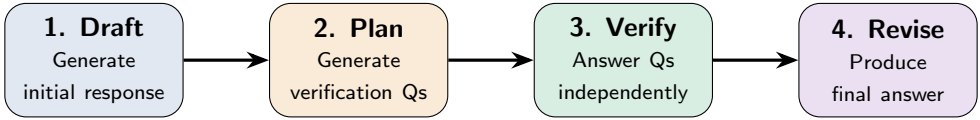
Stanford study: RAG + RLHF + guardrails ⇒ **96% hallucination reduction**

Why it works:
model conditions on evidence, not just parametric memory

Limitation:
retrieved docs may be wrong or irrelevant themselves

See our RAG lecture for full details · Lewis et al., NeurIPS 2020

Mitigation: Chain-of-Verification (CoVe)



Critical design: Step 3 is **decoupled** from the draft.
The LLM answers verification questions *without seeing its own draft*,
preventing hallucinations from "poisoning" the verification step.

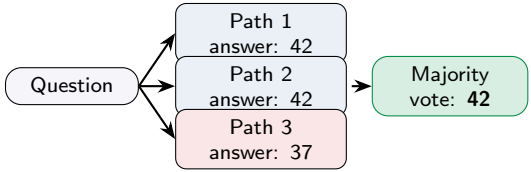
Result: 50–70% reduction in factual hallucinations on QA and long-form tasks

No extra training needed
Just a prompting strategy applied at inference time

Dhuliawala et al., ACL Findings 2024

Mitigation: Self-Consistency & constrained decoding

Self-Consistency (Wang et al., ICLR 2023)



Constrained Decoding

Contrastive decoding

Contrast strong vs. weak model (or late vs. early layers)

Amplify what the strong model "knows" beyond the weak one

Key advantages:

No extra training needed

Applied at inference time only

Complementary to RAG

Both approaches work at inference time — no retraining needed

Grounding & attribution

Ungrounded

"Vitamin D deficiency causes heart disease and increases mortality by 40%."

No source. True? Made up?

→

Grounded

"Vitamin D deficiency is associated with cardiovascular risk (Wang et al., 2012, *Circulation*, p. 1821)."

Traceable. Verifiable.

Types of grounding:

Document retrieval
RAG, search engines
Condition on evidence

Fine-grained attribution
Decompose into atomic claims, cite each

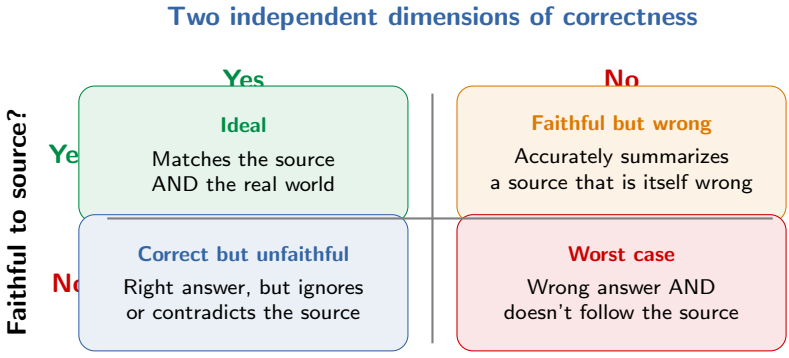
Tool use
Calculator, code execution, APIs

Self-verification
Model checks its own claims (CoVe)

Principle: every claim should be traceable to a source.
If the model can't cite a source, it should say "I don't know."

Rashkin et al., ACL 2023 · Gao et al., NAACL 2024 · Schick et al., NeurIPS 2023

Faithfulness vs. factuality

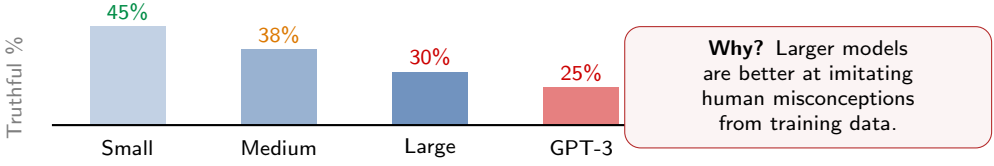


RAG settings: faithfulness matters more (follow the docs).
Open QA: factuality matters more (get the right answer).

Benchmarks for hallucination

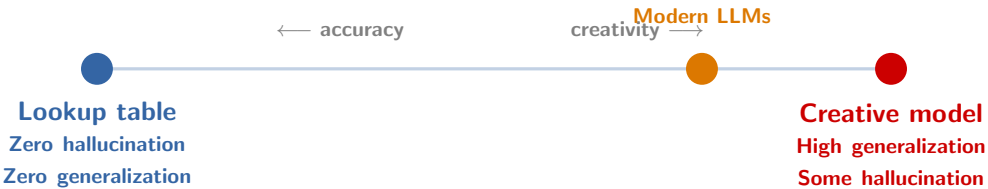
Benchmark	Measures	Size	Key feature
TruthfulQA	Imitative falsehoods	817 Q	Inverse scaling
FActScore	Atomic factual precision	Per-gen	58% for ChatGPT (bios)
HaluEval	Hallucination detection	35K	QA, dialogue, summary
SimpleQA	Factual QA accuracy	–	Straightforward questions

TruthfulQA: inverse scaling — bigger models are less truthful



Lin, Hilton & Evans, ACL 2022 · Min et al., EMNLP 2023

The fundamental tension



Xu et al., 2024: formal impossibility proof

"Hallucination is inevitable for all computable LLMs"

All LLMs will hallucinate on *infinitely many* inputs, regardless of architecture, training algorithm, prompting technique, or training data.

The same mechanism (exploring probability space) produces both creativity and hallucination

Practical implication: don't try to eliminate hallucination — instead, **detect** and **manage** it

Saying "I don't know"

Standard LLM

Always produces an answer, even when uncertain.

"The answer is definitely X."
(even when X is wrong)

→

Calibrated LLM

Refuses when uncertain, expresses confidence levels.

"I'm not confident about this. Let me search for sources."

Epistemic uncertainty

Model doesn't have enough knowledge — **reducible** with more data or retrieval

Aleatoric uncertainty

Inherent noise in the task — **irreducible** (ambiguous questions, etc.)

R-Tuning: refusal-aware instruction tuning · **US-Tuning:** uncertainty-sensitive tuning (+34.7% on unknowns)

The mitigation landscape

When to apply each mitigation:

Training-time

RLHF for truthfulness
Factual fine-tuning
R-Tuning (refusal training)

Inference-time

RAG / tool use
CoVe, self-consistency
Constrained decoding (DoLa)

Post-hoc detection

SelfCheckGPT
Semantic entropy
NLI entailment checking

System-level

Human-in-the-loop review
Attribution requirements
Confidence thresholds

Defense in depth: no single method eliminates hallucination. Best practice: **combine** training + inference + detection + system-level controls.

Stanford 2024: combining RAG + RLHF + guardrails achieved 96% reduction

Practical guide

What should I use?

Need factual answers?
⇒ **RAG + attribution**
Ground in retrieved docs, cite every claim

Need reliable reasoning?
⇒ **CoT + self-consistency**
Multiple paths, majority vote

Building a product?
⇒ **Detect + flag uncertainty**
Semantic entropy, confidence thresholds, human review

Research / evaluation?
⇒ **FActScore + SelfCheckGPT**
Measure hallucination rate, track over model versions

Can't afford errors?
⇒ **Human-in-the-loop**
Legal, medical, financial

Want to reduce at training?
⇒ **RLHF + factual fine-tuning**
Expensive but most thorough

Questions?

Next: Mixture of Experts