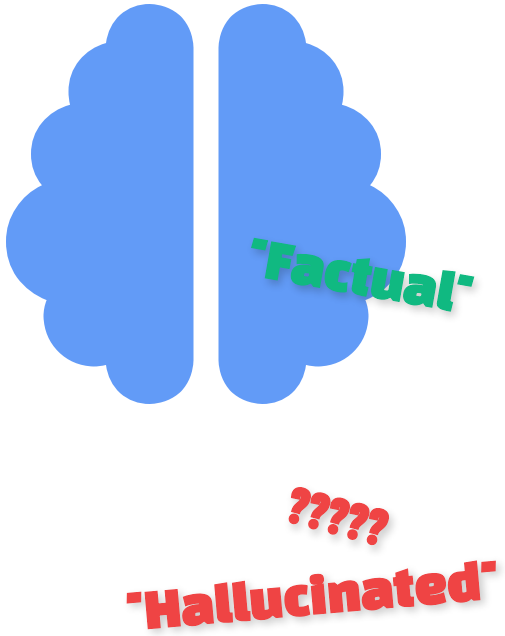
Understanding Hallucinations in LLMs

Detecting, Evaluating, and Preventing False Outputs in Large Language Models



> **"No Time for Hallucinations"**
>
> When AI confidently generates false information as if it were true

What Are Hallucinations?

Hallucinations occur when LLMs generate content that is:

Factually Incorrect | **Made-up Information** | **Misleading** | **Unfaithful to Source**

Factual Information

Source: "The Eiffel Tower was built in 1889 for the World's Fair."
LLM Output: "The Eiffel Tower, completed in 1889, was built as the entrance arch for the World's Fair in Paris."

Intrinsic Hallucination

Source: "The Eiffel Tower was built in 1889 for the World's Fair."
LLM Output: "The Eiffel Tower was built in **1878** by **Thomas Edison** as a radio broadcasting tower."

Completely contradicts facts in training data

Extrinsic Hallucination

Source: "The Eiffel Tower was built in 1889 for the World's Fair."
LLM Output: "The Eiffel Tower was built in 1889 for the World's Fair. **It was originally painted red before being repainted in its signature brownish-gray color in 1906.**"

Adds plausible but unsourced details

Why it matters: Hallucinations undermine trust, create legal/ethical risks, and can lead to harmful decision-making in critical applications.

Detection Methods

NLI-Based Methods

Uses **Natural Language Inference** to determine if document content **entails** or **contradicts** generated summary

> *"If a summary statement isn't entailed by any part of the source document, it's likely a hallucination."*

Examples: SummaC, TrueTeacher, MNLI models

Best Practice: Compare at sentence-level granularity (not whole documents)

Effectiveness: 75%

QA-Based Methods

Generate **questions** from document/summary and check if **answers match** when querying the other

> *"If answers from the document and summary differ for the same questions, inconsistencies exist."*

Examples: QuestEval, QAFactEval, SummaQA

Recall-Oriented
Q's from source → A's from summary

Precision-Oriented
Q's from summary → A's from source

Effectiveness: 80%

Reference-Based Metrics

Compare generated output with **gold standard references** using similarity metrics

> *"Significant deviation from trusted references may indicate hallucination."*

ROUGE | METEOR | BERTScore | MoverScore

Limitation: Requires high-quality references, which are often unavailable

Effectiveness:

LLM-as-Judge

Use a **larger LLM** to evaluate outputs for factual consistency and hallucinations

> *"One LLM can critique another's output for accuracy using chain-of-thought reasoning."*

Examples: G-Eval, SelfCheckGPT, specialized reward models

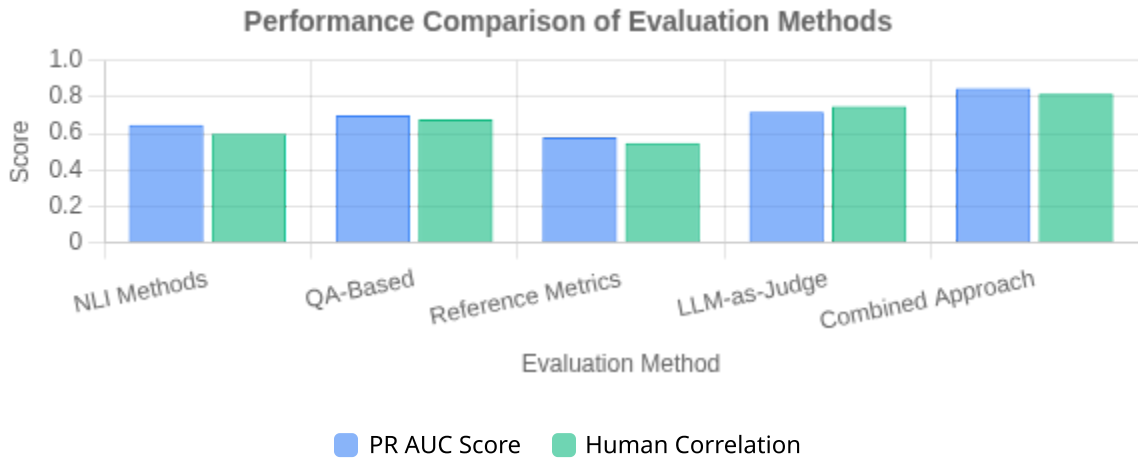
Best Practice:

- Use judge models much larger than evaluated model
- Fine-tune specifically for evaluation tasks

Effectiveness: 85%

Combining multiple detection methods creates more robust hallucination detection systems than any single approach alone.

Evaluation Techniques



Key Insights

Balanced accuracy for detecting inconsistency on thoroughly studied datasets ranges from **60-75%** in state-of-the-art approaches.

Combining multiple evaluation methods yields the most reliable hallucination detection systems.

Evaluation Process

Benchmark Testing → Human Evaluation → Production Monitoring

Evaluation must be continuous and adapt to domain-specific needs

Reference-based vs Reference-free

Reference-based

- ROUGE, METEOR, BERTScore
- Traditional & widely used
- Requires gold-standard references

> *"Reference summaries often score poorly on relevance, consistency, and coherence"*

Reference-free

- ROUGE-C, direct source comparison
- No reference needed
- Better for new domains

> *"More practical for real-world applications with domain-specific data"*

Pre-finetuning on out-of-domain data improved PR AUC from 0.69 to 0.85 (23% increase)

Advanced Techniques

LLM-as-Judge

Using larger LLMs to evaluate outputs for factuality

Effectiveness: **85%**

Preference-Based

Human feedback & reward modeling approaches

Effectiveness: **78%**

Sampling-Based

Testing consistency across multiple generations

Effectiveness: **72%**

QA Approach

Question generation & answer verification

Effectiveness: **80%**

Challenge:
Most evaluation methods report correlations with human annotations rather than interpretable metrics like precision/recall

Pragmatic Evaluation Approach

1. Start with reference-based metrics if available
2. Apply NLI models for consistency evaluation
3. Use sampling-based verification approaches
4. Develop a reward model with human preferences

Remember: Evaluation should be iterative and continuously improved as your application evolves

Real-World Hallucination Examples

Entity Swapping

Source: *"Vehicles and pedestrians will now embark and disembark the Cowes ferry separately following Maritime and Coastguard Agency guidance."*

Generated: *"A **new service** on the Isle of Wight's chain ferry has been **launched** following a complaint from a resident."*

Reality: *The source discusses **safety measures** on an **existing service**, not a new launch.*

HIGH IMPACT — Completely changes the meaning of the information

Factual Fabrication

Source: *"Wendy Jane Crewson was born in Hamilton, Ontario, the daughter of June Doreen and Robert Binnie Crewson."*

Generated: *"Wendy Jane Crewson (born **May 9, 1956**) is a Canadian actress and producer."*

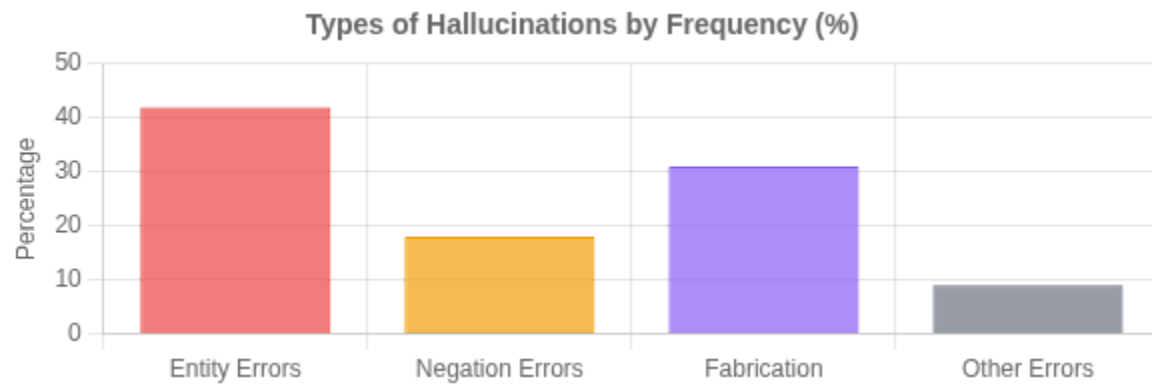
HIGH IMPACT — Adds specific details not present in the source

Negation Flips

Source: *"Studies show the treatment is effective for 60% of patients with condition X."*

Generated: *"The treatment **is not effective** for a significant portion of patients with condition X."*

MEDIUM IMPACT — Reverses the core claim while maintaining partial truth



Business Impact

- User trust erosion when incorrect information is presented confidently
- Legal/compliance risks when hallucinated content violates regulations
- Effective detection can provide 23% performance improvement

30%
CNN/DailyMail summaries contain hallucinations

92%
XSum summaries contain faithfulness errors

43%
Average error rate across evaluated datasets

Source: Pagnoni et al. (2021), Kryciski et al. (2020)

> *Even well-trained LLMs struggle with hallucinations when dealing with unfamiliar domains or ambiguous contexts*

Mitigating Hallucinations

Out-of-Domain Finetuning

Pre-finetune on related domains before task-specific finetuning

23% IMPROVEMENT

NLI-Based Detection

Use entailment models to verify factual consistency

SEMANTIC VALIDATION

QA-Based Methods

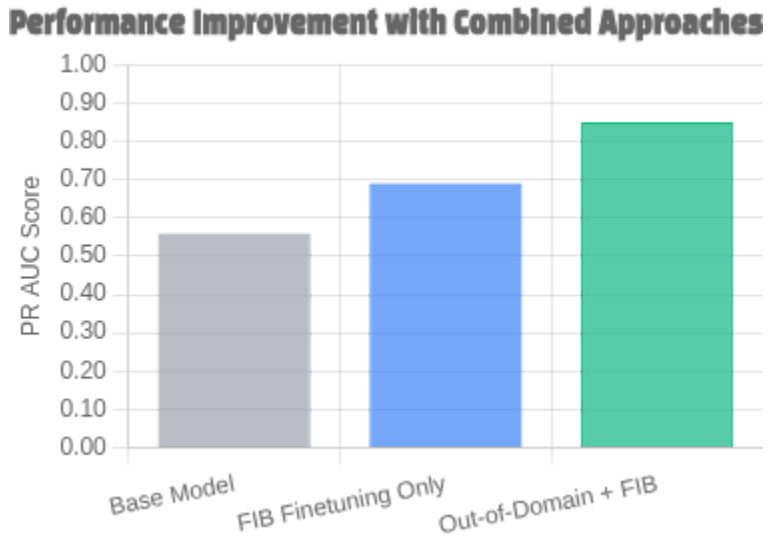
Generate questions from context to validate summary claims

PRECISION/RECALL BALANCE

LLM-as-Judge

Use larger models to evaluate outputs for factual consistency

HUMAN-ALIGNED EVALUATION



0.85
PR AUC After Combined Approaches

50x
Recall Improvement at 0.8 Threshold

Key Success Factors

- Better separation of probability distributions
- Continuous evaluation across domains
- Combined methods outperform single approaches

Complete Mitigation Pipeline

1. Data Preparation
Identify out-of-domain & in-domain datasets with factual consistency labels

2. Model Pre-finetuning
Finetune base model on out-of-domain data first (e.g., Wikipedia summaries)

3. Task-specific Finetuning
Further finetune on in-domain data (e.g., news summaries)

4. Multi-method Evaluation
Apply NLI, QA, and reference-based metrics

5. Threshold Optimization
Set production thresholds based on precision/recall requirements

6. Continuous Monitoring
Apply LLM-as-judge to detect emerging hallucination patterns

Research Insight

"Bootstrapping with Wikipedia summaries improved factual inconsistency classification in news summaries, even though the former is out-of-domain."

— Yan, Ziyou (2023)

Effective hallucination mitigation requires a comprehensive approach across the entire ML lifecycle

Conclusion & Future Directions

Key Insights

Understanding Hallucinations

Hallucinations occur when LLMs generate content that is unfaithful to source material or contains fabricated facts

Multiple Detection Approaches

NLI, QA-based, and reference metrics each provide unique insights into different aspects of hallucination

Evaluation is Multi-dimensional

Effective evaluation combines controlled benchmarks, human judgment, and field testing

Effectiveness of Transfer Learning

Out-of-domain finetuning significantly boosts performance on in-domain hallucination detection

Future Research Directions

Architectural Improvements

Self-verification mechanisms built directly into model architectures

Multi-modal Verification

Cross-referencing information across different modalities

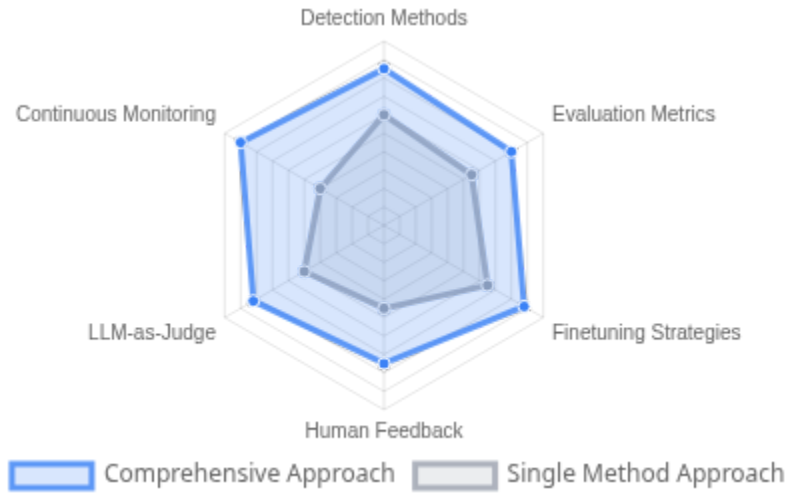
Uncertainty Quantification

Better calibration of model confidence in generated content

Human-AI Collaboration

Frameworks that optimize human oversight of hallucination-prone content

Hallucination Management Cycle



Key Takeaways

Hallucination Detection is Multi-faceted

Combine NLI-based, QA-based, and reference-free approaches for robust detection

Continuous Evaluation is Critical

"Evaluation is not a one-time job... this process will be continuous"

Domain-Specific Benchmarks Matter

Models that excel on public benchmarks often fail on domain-specific evaluations

Transfer Learning Boosts Performance

23% improvement in PR AUC with out-of-domain + in-domain finetuning strategy

> *"The evaluation pipeline should be flexible, automated, and able to scale with production needs"*

Mitigating hallucinations is an ongoing challenge requiring continuous innovation and evaluation