1. Forms of Hallucinations

Hallucinations are not all the same.[1] They are typically categorized based on **source conflict** (Intrinsic vs.[2] Extrinsic) or **type of error** (Factuality vs. Faithfulness).

+1

Form	Definition	Example
Intrinsic Hallucination	The model's output **directly contradicts** the provided source content or context. The answer is "internally" inconsistent with the input.	**Input:** "The Eiffel Tower is in Paris." **Output:** "The Eiffel Tower, located in London..."
Extrinsic Hallucination	The model generates information that is **not present** in the source. It might be factually correct in the real world, but it is "hallucinated" relative to the strict context provided (unverifiable).	**Input:** "Steve Jobs founded Apple." **Output:** "Steve Jobs founded Apple and loved sushi." (If the input didn't mention sushi, this is extrinsic).
Factuality Hallucination	The output contradicts established **real-world knowledge**.	"The first person to walk on Mars was Neil Armstrong."
Faithfulness Hallucination	The output diverges from the **user's instruction** or the **logic** of the	**Instruction:** "Translate to Spanish."

	prompt, even if the facts are technically correct.	**Output:** (Translates to French).

---

2. Hallucination Control at Various Levels

Controlling hallucinations requires a "defense-in-depth" strategy, applying controls at the Data, Model, Retrieval (RAG), and Prompting levels.

Level 1: The Data Level (Pre-Training & Knowledge Base)

- **Curated Data Cleaning:** Remove duplicate, contradictory, or low-quality data from the training set.[3] "Garbage in, garbage out" is the primary driver of factuality errors.
- **Knowledge Graph Integration:** Structure unstructured data into Knowledge Graphs (KGs). This forces the model to traverse defined entities and relationships (Subject $\rightarrow$ Predicate $\rightarrow$ Object) rather than probabilistically guessing the next word.

Level 2: The Model Level (Fine-Tuning)

- **Domain Adaptation (SFT):** Fine-tune the model on domain-specific high-quality datasets (e.g., medical or legal papers) to align its internal weights with specialized knowledge.[4]
- **RLHF (Reinforcement Learning from Human Feedback):** Train a Reward Model to penalize hallucinations.[5] If the model guesses and gets it wrong, it receives a negative reward, teaching it to be "honest" and refuse to answer when unsure.
- **Rejection Sampling / Negative Training:** Train the model specifically on examples where it *should* say "I don't know" rather than making up an answer.[6]

Level 3: The Retrieval Level (RAG Systems)

Retrieval-Augmented Generation (RAG) is the most effective architectural pattern for reducing hallucinations by grounding the model in external evidence.[7]

- **Advanced Chunking:** Avoid breaking text in the middle of a sentence. Use **Semantic Chunking** to keep related ideas together so the model doesn't lose context.
- **Strict Context Grounding:** Force the model to answer *only* using the retrieved chunks.[8]

    - *Technique:* "Answer solely based on the provided context.[9] If the answer is not there, state that you do not know."
- **Citation/Attribution:** Require the model to cite the specific document ID or paragraph number for every claim it makes. If it cannot find a citation, it is likely hallucinating.

Level 4: The Prompting Level (Inference)

- **Temperature = 0:** Set the temperature parameter to 0 to make the model deterministic and less "creative."[10]
- **Chain of Thought (CoT):** Ask the model to "think step-by-step."[11] This forces the model to lay out its logic before committing to a final answer, reducing logical inconsistency.[12]
- +1
- **Chain of Verification (CoVe):** A multi-step prompting method:
    1. Model generates a draft response.
    2. Model generates "verification questions" to fact-check its own draft.[13]
    3. Model answers those questions independently to verify facts.
    4. Model generates the final corrected answer.
- **Self-Consistency:** Generate 5-10 different answers for the same prompt and pick the "majority vote" (the answer that appears most often). Hallucinations tend to be random; truth tends to be consistent.

Level 5: The Post-Processing Level (Guardrails)

- **Hallucination Detection Models:** Use a smaller, specialized model (like a Natural Language Inference or NLI model) to check if the LLM's output is actually supported by the source text.
- **Confidence Scores (Log-probs):** Check the "token probability" (log-probs) of the generated answer. If the model's confidence for a specific named entity (e.g., a person's name or date) is low, flag the response for human review.