

MCQ Question · mcqquestion.com

Large Language Models (LLMs)

Free · Share

Large Language Models (LLMs)

An LLM does not look up an answer. It calculates the odds of every possible next token, picks one, and repeats. That one loop explains why LLMs sound fluent, why they can be biased, and why they sometimes state falsehoods with confidence. UPSC tested exactly these ideas in 2026. This lesson explains the mechanics step by step, then covers India’s sovereign LLM push with figures dated up to August 2026.

Must know: tokens, the next-word loop and attention

1 What is a Large Language Model, and why is it called "large"?

An LLM is a neural network trained on huge amounts of text. Its basic skill is predicting the next token. GPT-3, described in a 2020 paper, is a clear early example. It is an autoregressive language model. Autoregressive means each new token is produced from the tokens before it.



An NVIDIA H100 accelerator card. Photo: Geekerwan, CC BY 3.0, via Wikimedia Commons. Resized.

What does “large” actually measure?

“Large” points to three things that grow together.

- **Parameters:** the adjustable numbers inside the model. GPT-3 had **175 billion**.
- **Training data:** GPT-3 was trained on a total of **300 billion tokens**.
- **Compute:** the processing power used during training.

A 2020 study found that the model’s error falls smoothly as model size, data size and compute all increase. Researchers call this a power law. The same study says model size, data and compute must be scaled up together.

How big have models become?

In 2024, Meta described Llama 3, whose largest model is a dense transformer with **405 billion parameters**. It has a context window of up to **128K tokens**. The context window is how much text the model can consider at once.

What goes into the training text?

GPT-3's mix was mostly web text. Filtered Common Crawl made up 60% of the training mix by weight. The rest was WebText2, two book collections and an online encyclopedia. Whatever sits in this text, good or bad, can shape the model.

Is a bigger model always better?

No. A 2022 study trained Chinchilla, a 70-billion-parameter model, on four times more data than Gopher used. Chinchilla beat Gopher (280 billion) and GPT-3 (175 billion) on many tasks. The lesson is that model size and training data must grow together.

> **Quick check 1.** A model with more parameters always beats a smaller model, whatever data each was trained on. *(True or False?)* See answer ↓

Sources: arxiv.org

2 How does text become tokens?

First, a tokenizer cuts the text into small pieces called tokens. The transformer paper then uses learned embeddings to convert input tokens to vectors of numbers.

Is a token the same as a word?

No. A token can be a whole word. It can also be part of a word. A 2015 research paper showed that rare and unknown words can be written as sequences of **subword units**. The paper adapted a compression method, byte-pair encoding, to split words this way.

Why do tokens matter for Indian languages?

Sarvam AI says its tokenizer is optimised for efficient tokenization across all 22 scheduled Indian languages, spanning 12 scripts. The company says this directly reduces the cost and latency of serving in Indian languages.

> **Quick check 2.** A token can be a piece of a word, not only a whole word. *(True or False?)* See answer ↓

Sources: arxiv.org; sarvam.ai

3 How does an LLM choose the next word?

The transformer paper converts its output into predicted next-token probabilities. It calls the model auto-regressive, so each step uses the symbols already produced. A decoding rule picks which token to use.

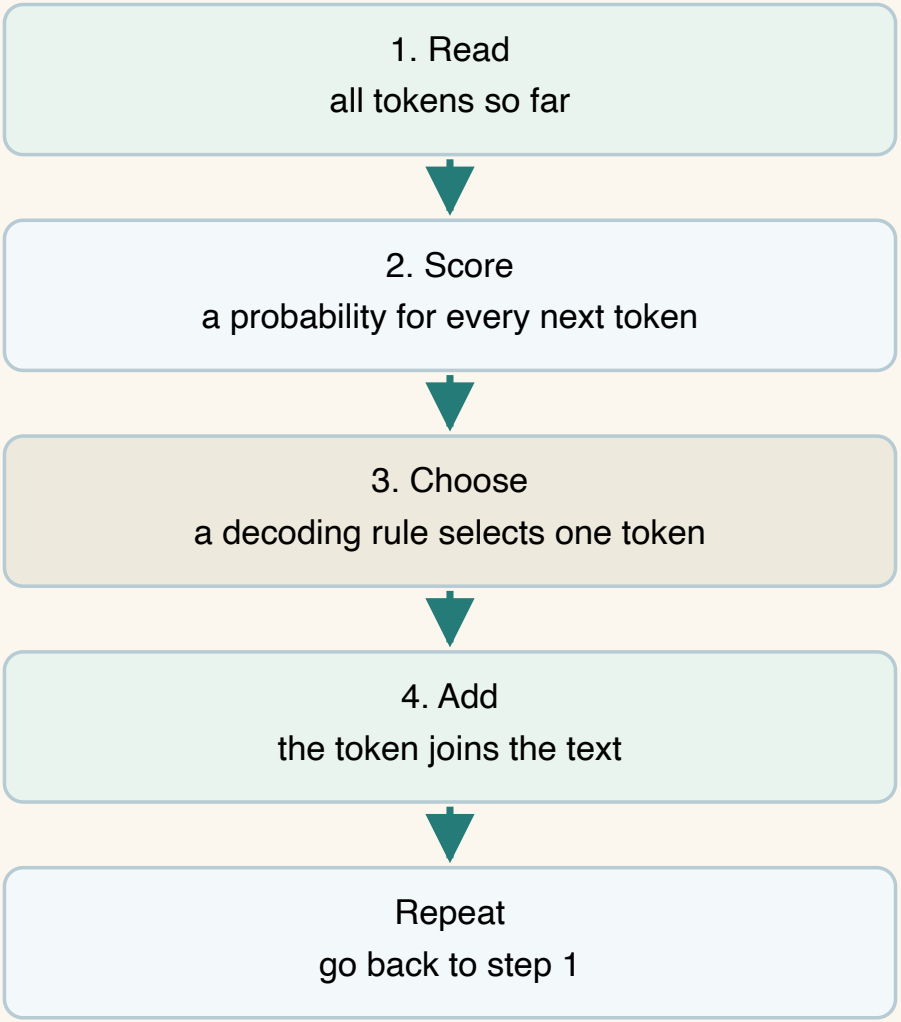
What are the steps?

1. The model reads all tokens so far.
2. It calculates a probability for every possible next token.
3. A decoding rule selects one token.
4. The token is added, and the model repeats from step 1.

Does the model always pick the highest-probability token?

Not always. Choosing the single most likely token every time is called greedy decoding. A 2019 study found that using likelihood as the decoding rule gives text that is bland and strangely repetitive. The authors proposed **nucleus sampling**. It draws a token at random from the top slice of the distribution and cuts off the unreliable tail.

So “picks the highest-probability token” describes one rule. Nucleus sampling is an alternative rule.



The model writes one token at a time. It repeats this loop for every token. Original explanatory diagram.

Sources: arxiv.org

4 What does the transformer's self-attention do?

The **transformer** is the architecture behind GPT-3 and Llama 3. It was introduced in the 2017 paper “Attention Is All You Need”. The paper proposed a network based only on attention. It dropped recurrence and convolutions entirely.

What is self-attention?

Self-attention is a mechanism that relates different positions in one sequence. It helps compute a representation of that sequence. In the paper’s words, an attention function maps a query and key-value pairs to an output. The output is a weighted sum of the values.

The weight given to each value comes from a compatibility function of the query with the corresponding key.

Why did transformers win?

The 2017 paper reported two gains.

- The model was more parallelizable. Many computations could run at the same time.
- It needed much less time to train. One result came after 3.5 days on eight GPUs.

Do transformers write the whole answer at once?

No. The paper calls the model more parallelizable, so training is faster. Generation is still step by step. The paper describes the model as auto-regressive. It consumes the symbols it has already produced when it generates the next one.

Sources: arxiv.org

Good to know: training, bias and hallucination

5 How is an LLM trained, and then made more helpful?

Training starts with pre-training on raw text. Then people shape the behaviour, as InstructGPT shows in three steps.

What happens in pre-training?

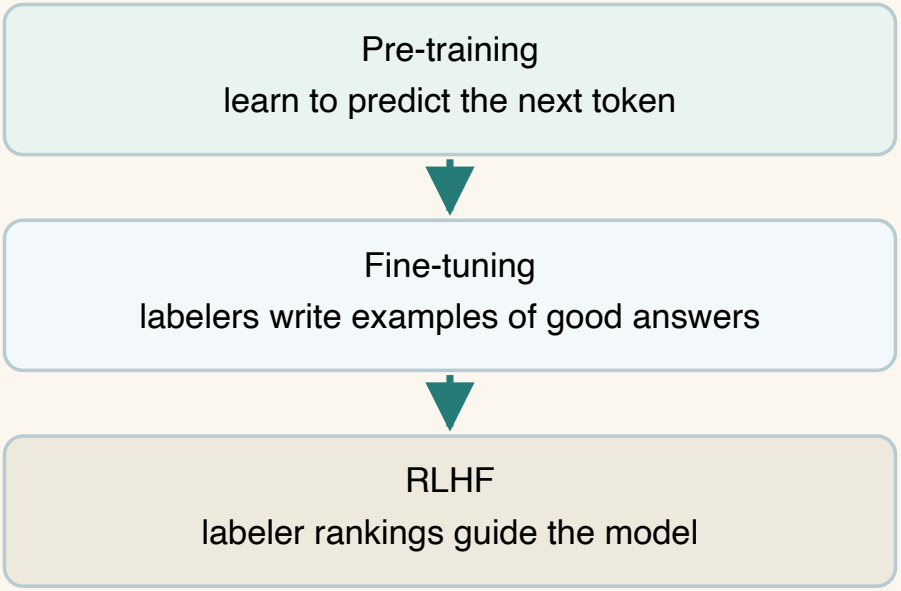
The model is shown text and asked to predict the next token. Its error is measured as a **loss**. The 2020 scaling study measures this loss as cross-entropy. An optimiser then adjusts the parameters to reduce the loss. GPT-3 used the Adam optimiser and clipped the gradient at 1.0. The Adam paper describes Adam as an algorithm for first-order gradient-based optimization.

What is fine-tuning with human feedback?

A pre-trained model can still be untruthful, toxic or unhelpful. A 2022 paper on InstructGPT describes a fix in steps.

1. Labelers wrote examples of good answers. These were used to fine-tune GPT-3.
2. Labelers ranked several model outputs.
3. Those rankings were used in reinforcement learning from human feedback, often called **RLHF**.

Human evaluators preferred outputs from the 1.3-billion-parameter InstructGPT over the 175-billion-parameter GPT-3. The paper also reported better truthfulness and less toxic output. It still said the models make simple mistakes.



Raw text comes first. Human examples and rankings come after. Original explanatory diagram.

> **Quick check 3.** Human feedback can be used to fine-tune an LLM after its first training stage. *(True or False?)* See answer ↓

Sources: arxiv.org

6 Why can LLM outputs be biased?

An LLM learns from its training data. It has no built-in sense of fairness. So patterns in the data can show up in the output.

What does the evidence say?

- The GPT-3 paper says biases in training data may lead models to generate stereotyped or prejudiced content.

- A 2016 research paper trained a word-embedding model on ordinary web text. The model reproduced human-like biases found in psychology tests.
- The US National Institute of Standards and Technology (NIST) lists “harmful bias or homogenization” as a generative-AI risk. Its July 2024 profile includes performance gaps between groups or languages, possibly due to non-representative training data.

What is India doing about it?

The IndiaAI Mission has a Safe and Trusted AI pillar. PIB says it focuses on bias mitigation, machine unlearning, privacy-preserving designs, algorithm auditing and deepfake detection. PIB reported 13 projects selected under it.

What is the exam point?

A claim that LLMs “produce unbiased outputs” is false. Bias is a documented risk, and researchers work to reduce it.

> **Quick check 4.** Because an LLM only predicts statistically likely text, its outputs are free of bias. *(True or False?)* See answer ↓

Sources: arxiv.org; nvlpubs.nist.gov; pib.gov.in

7 What is hallucination, and can it be fixed?

Hallucination is fluent, confident text that is false. NIST uses the term **confabulation**. It defines it as confidently stated but erroneous or false content, known colloquially as hallucinations or fabrications.

Why does it happen?

- A 2022 survey says deep-learning-based text generation is prone to hallucinate unintended text.
- A September 2025 research paper argues that models hallucinate because training and testing reward guessing over admitting uncertainty. It compares this to a student guessing on a hard exam.

What reduces it?

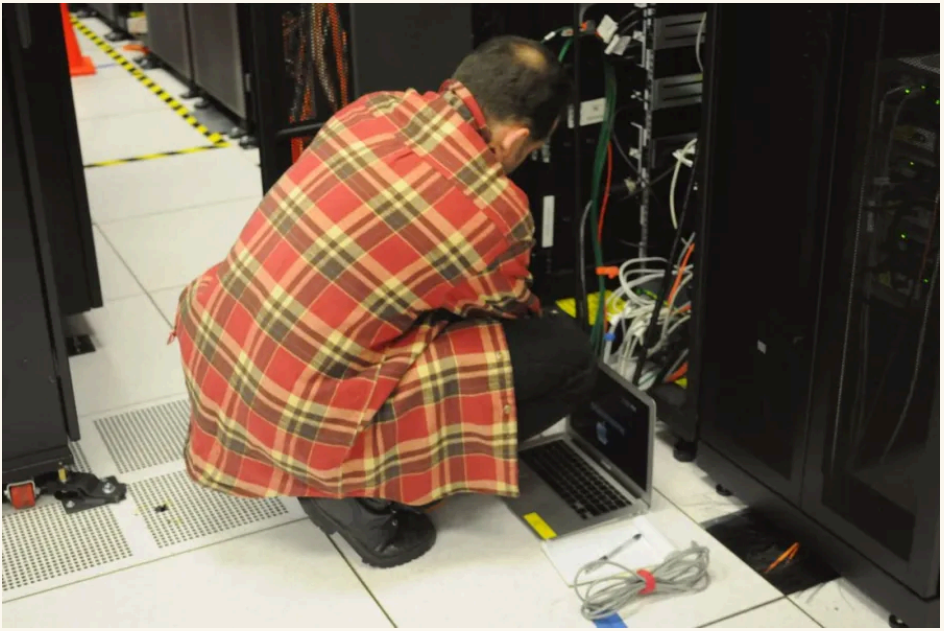
- **Retrieval-augmented generation (RAG).** A 2020 paper joined a language model to an external memory. The authors say plain models struggle to show provenance and to update knowledge. They say a model with access to explicit memory can overcome this issue.
- **Human feedback.** InstructGPT improved truthfulness but still made simple mistakes.

Sources: nvlpubs.nist.gov; arxiv.org

Great to know: India's models and the 2026 UPSC question

8 What is India building, as of August 2026?

India wants LLMs trained on Indian languages and data. The effort sits inside the IndiaAI Mission. The Summit story is in SciTech0139, AI Impact Summit 2026. The research centres and skilling labs are in SciTech0044, AI Centres of Excellence in India. Both are linked under Related lessons.



A technician works beside server racks at the NERSC data centre in the United States. Compute at this scale sits in rooms like this. Image credit: Derrick Coetzee, released under CC0, source Wikimedia Commons. Resized.

What is the IndiaAI Mission?

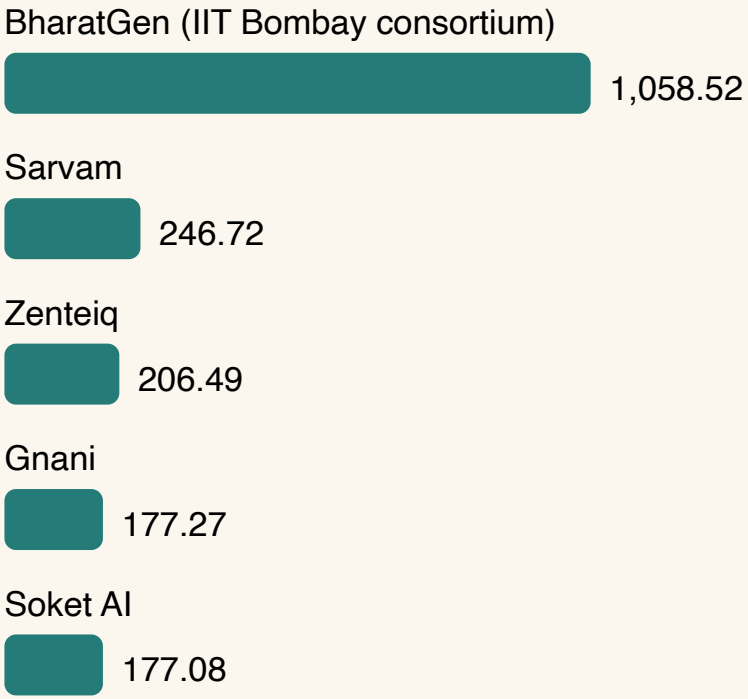
On 7 March 2024 the Cabinet approved the mission with an outlay of **Rs 10,371.92 crore**. The IndiaAI Independent Business Division under the Digital India Corporation runs it. It has seven components.

- Compute capacity
- Innovation Centre
- Datasets platform
- Application development
- FutureSkills
- Startup financing
- Safe and Trusted AI

Who is building foundation models, and with how much support?

PIB's 13 February 2026 release says twelve organisations were selected under the Innovation Centre (Foundation Models) pillar. Support covers actual compute costs. An extra 25% of compute spending is set aside for datasets and staff. The five largest totals in PIB's annexure are below.

Organisation	Total support (Rs crore)
BharatGen (IIT Bombay consortium)	1,058.52
Sarvam	246.72
Zenteiq	206.49
Gnani	177.27
Soket AI	177.08



Total support in Rs crore. BharatGen is far ahead of the rest. Original explanatory diagram.

BharatGen's total is Rs 990.92 crore of compute plus Rs 67.60 crore of other costs. PIB (25 November 2025) says BharatGen is supported under the NM-ICPS of the Department of Science and Technology. It says Rs 235 crore is being channelled through the Technology Innovation Hub at IIT Bombay.

What are the two flagship models?

- **Sarvam 30B and 105B.** Sarvam says both are reasoning models trained from scratch, in India, on compute from the IndiaAI Mission. Both use a Mixture-of-Experts transformer. Only some experts work on each token. The 30B model has 2.4 billion active parameters and was trained on 16 trillion tokens. The 105B model was trained on 12

trillion tokens. Sarvam released them as open source under Apache 2.0.

- **BharatGen Param2 17B.** BharatGen describes it as a 17-billion-parameter Mixture-of-Experts model that activates 2.4 billion parameters per token. It was trained on about 22 trillion tokens. It supports English and all 22 scheduled languages. PIB (25 November 2025) says the BharatGen consortium is led by IIT Bombay.

What is the latest progress?

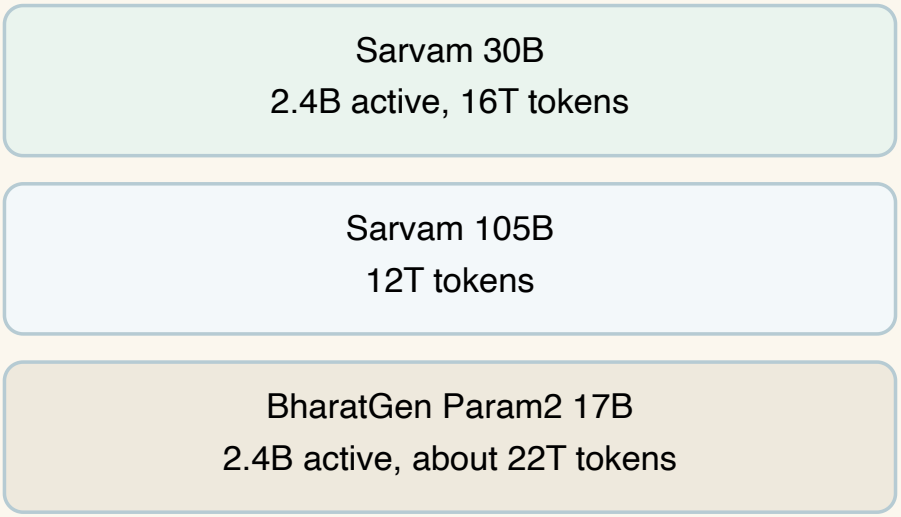
- By August 2026 the Government had selected **20** indigenous foundation-model proposals from 506 applications. Twelve are large multimodal models and eight are small language models.
- Shared compute had passed **45,000 GPUs** by June 2026.
- AIKosh, the national datasets platform, held over 14,000 datasets and 331 AI models by July 2026.

What are India's governance rules?

MeitY released the India AI Governance Guidelines on 5 November 2025. They rest on seven Sutras.

1. Trust
2. People First
3. Innovation over Restraint
4. Fairness and Equity
5. Accountability
6. Understandable by Design
7. Safety, Resilience and Sustainability

A committee chaired by Prof. Balaraman Ravindran of IIT Madras drafted them. They favour a risk-based approach. They propose an AI Governance Group, a Technology and Policy Expert Committee and an AI Safety Institute. MeitY's Secretary said the focus is on using existing legislation wherever possible.



Mixture-of-Experts models use only some of their parameters for each token. Original explanatory diagram.

> **Quick check 5.** Of the twelve foundation-model organisations in PIB's annexure, Sarvam received the largest total support. *(True or False?)* See answer ↓

Sources: pib.gov.in; sarvam.ai; bharatgen.com

9 How did UPSC test LLMs in 2026, and how do you answer?

The Civil Services (Preliminary) Examination 2026 was held on 24 May 2026. General Studies Paper I, Set A, Question 42 read as follows.

> Which of the following statements with regard to Large Language Models (LLMs) used in machine learning is/are correct?
> 1. LLMs assign probabilities to the next possible words and then pick the one with the highest probability.
> 2. LLMs process data through mathematical optimization to minimise prediction errors.
> 3. LLMs produce unbiased outputs.
> Select the answer using the code given below :
> (a) 1 only
> (b) 1 and 2 only
> (c) 2 and 3 only
> (d) 1, 2 and 3

How do you work it out?

- **Statement 2 is correct.** Training measures prediction error as a loss. An optimiser reduces it (section 5).
- **Statement 3 is false.** LLMs inherit bias from their data (section 6).
- **Statement 1 is the subtle one.** The model does assign probabilities to next tokens. “Pick the highest” is one decoding rule, and many systems sample (section 3).
- **Use elimination.** Statement 3 is false, so options (c) and (d) go. Statement 2 is true, so option (a) goes. Only (b), “1 and 2 only”, remains.

UPSC publishes its official key only after the final result. Statement 1 is the weak one. Even so, only (b) survives elimination.

What are the common traps?

- Calling LLM output “unbiased” or “verified”.
- Mixing the number of parameters with the number of training tokens.

- Mixing BharatGen's Rs 1,058.52 crore with the IndiaAI Mission's Rs 10,371.92 crore.
- Mixing 12 organisations with 20 selected proposals.

Sources: upsc.gov.in; arxiv.org; nvlpubs.nist.gov

Review the feedback after each answer. The understanding-check score lasts for this page session.

Reset the checks when you want another attempt. This score is separate from the Native Practice score below.

Questions to think through

Why does an LLM sound sure even when it is wrong?

The model is trained to produce likely text, not to check facts. A September 2025 paper adds that testing methods often reward a guess over an honest "I am not sure". So a confident wrong answer is a normal failure, not a rare glitch.

Is "pick the highest-probability word" always the best rule?

No. A 2019 study found that always choosing the likeliest token gives bland, repetitive text. Sampling from the top slice of the distribution gave text closer to human writing. Greedy decoding is one option, not the only one.

Why does India want its own LLMs when foreign models exist?

PIB says the aim is sovereign capability trained on Indian datasets and languages. Sarvam adds that its tokenizer is built for all 22 scheduled languages. A model built for those languages can serve them at lower cost and delay.

Can a smaller model beat a bigger one?

Yes, in two documented ways. Chinchilla, at 70 billion parameters, beat larger models after training on more data. InstructGPT, at 1.3 billion parameters, was preferred by human evaluators over 175-billion-parameter GPT-3. Size alone does not decide quality.

Practice MCQs

1. What is a "token" in the context of how a Large Language Model processes text?

A. A small chunk of text, roughly a word or part of a word

B. A single training example drawn from the dataset

C. A unit of GPU memory allocated during training

D. A cryptographic key used to authenticate API requests

See answer ↓

2. Modern Large Language Models are built on an architecture introduced in the 2017 paper "Attention Is All You Need". What is this architecture called?

A. Transformer

B. Recurrent neural network

C. Generative adversarial network

D. Convolutional neural network

See answer ↓

3. During training, an LLM's prediction error is measured as a loss.

Which kind of method then adjusts its parameters to reduce that loss?

A. Gradient-based optimisation, such as the Adam optimiser

B. Retrieval of answers from an external database

C. Human labelling of every sentence in the corpus

D. Rule-based programming of grammar by linguists

See answer ↓

4. Consider the following statements about Large Language Models:

1. Biases in training data can lead an LLM to produce stereotyped content.

2. Hallucination means producing fluent, confident text that is false.

3. Fine-tuning with human feedback removes hallucination completely.

Which of the statements given above is/are correct?

- **A.** 1 only
- **B.** 1 and 3 only
- **C.** 2 and 3 only
- **D.** 1 and 2 only

See answer ↓

5. A PIB release of 13 February 2026 listed support under the IndiaAI Innovation Centre (Foundation Models) pillar.

Which organisation received the largest total support?

- **A.** Gnani, listed by PIB as Gnani AI
- **B.** Sarvam, maker of the Sarvam 30B and 105B models
- **C.** BharatGen, the IIT Bombay-led consortium
- **D.** Soket AI, also listed by PIB among the twelve

See answer ↓

6. As of August 2026, from 506 applications, how many indigenous foundation model proposals had the Government selected under the IndiaAI Mission?

A. 20

B. 12

C. 8

D. 506

See answer ↓

7. The Cabinet approved the IndiaAI Mission on 7 March 2024. What budget outlay did it approve?

A. Rs 1,058.52 crore

B. Rs 10,371.92 crore

C. Rs 246.72 crore

D. Rs 76,000 crore

See answer ↓

Answer Key

1. **A.** An LLM first splits text into tokens. A token is roughly a word or part of a word. Research on subword units showed that a rare word can be written as several smaller pieces. The model then calculates a probability for every possible next token. The other options describe a security key, GPU memory and a training example. None of these is a token. ↑ Q1
2. **A.** The 2017 paper proposed the Transformer. It is based only on attention and drops recurrence and convolutions. It trains faster because many computations run in parallel. GPT-3 uses transformer layers, and the largest Llama 3 model is a dense Transformer. Recurrent and convolutional networks were the older approach that the paper replaced. ↑ Q2
3. **A.** Pre-training measures error as a cross-entropy loss. A gradient-based optimiser then nudges the parameters to lower it. GPT-3 used the Adam optimiser and clipped the gradient at 1.0. The Adam paper describes Adam as an algorithm for first-order gradient-based optimization. Retrieval is a separate technique that helps at answer time, not in training. ↑ Q3
4. **D.** Statements 1 and 2 are correct. The GPT-3 paper warns that biased data can give stereotyped output. NIST defines confabulation as confidently stated but false content. Statement 3 is wrong. InstructGPT improved truthfulness with human feedback, but its authors say it still makes simple mistakes. ↑ Q4
5. **C.** BharatGen received Rs 1,058.52 crore, made of Rs 990.92 crore compute support and Rs 67.60 crore other costs. Sarvam received Rs 246.72 crore. Gnani received Rs 177.27 crore and Soket AI Rs 177.08 crore. Twelve organisations were selected in all. PIB of 25 November 2025 says BharatGen is supported under the NM-ICPS of the Department of Science and Technology. It says Rs 235 crore is channelled through the Technology Innovation Hub at IIT Bombay. ↑ Q5
6. **A.** PIB's factsheet of 13 August 2026 says the Government selected 20 proposals from 506 applications. Twelve are large multimodal models and eight are small language models. Do not confuse this with the twelve organisations that PIB listed in February 2026. Eight is only the count of small language models. ↑ Q6

7. **B.** The Cabinet approved Rs 10,371.92 crore for the IndiaAI Mission, to be run by the IndiaAI Independent Business Division under the Digital India Corporation. Rs 1,058.52 crore is BharatGen's support and Rs 246.72 crore is Sarvam's. Rs 76,000 crore is the Semicon 1.0 outlay, a separate semiconductor programme. ↑ Q7

Quick-check answers

1. **False.** Right to doubt it. Chinchilla, at 70 billion parameters, beat larger models after training on four times more data than Gopher used. ↑ 1
2. **True.** Tokenizers use subword units, so a rare word can be split into several tokens. ↑ 2
3. **True.** InstructGPT used labeler examples and then human rankings of outputs in reinforcement learning from human feedback (RLHF). ↑ 3
4. **False.** An LLM learns from its training data, so biases in that data can appear in its output. This is the false statement in the UPSC 2026 question. ↑ 4
5. **False.** BharatGen received the largest total, Rs 1,058.52 crore. Sarvam's total was Rs 246.72 crore. ↑ 5

Curated by Arsalan Zuberi · mcqquestion.com

Scan to read online

https://mcqquestion.com/articles/scitech0136-large-language-models-llms/