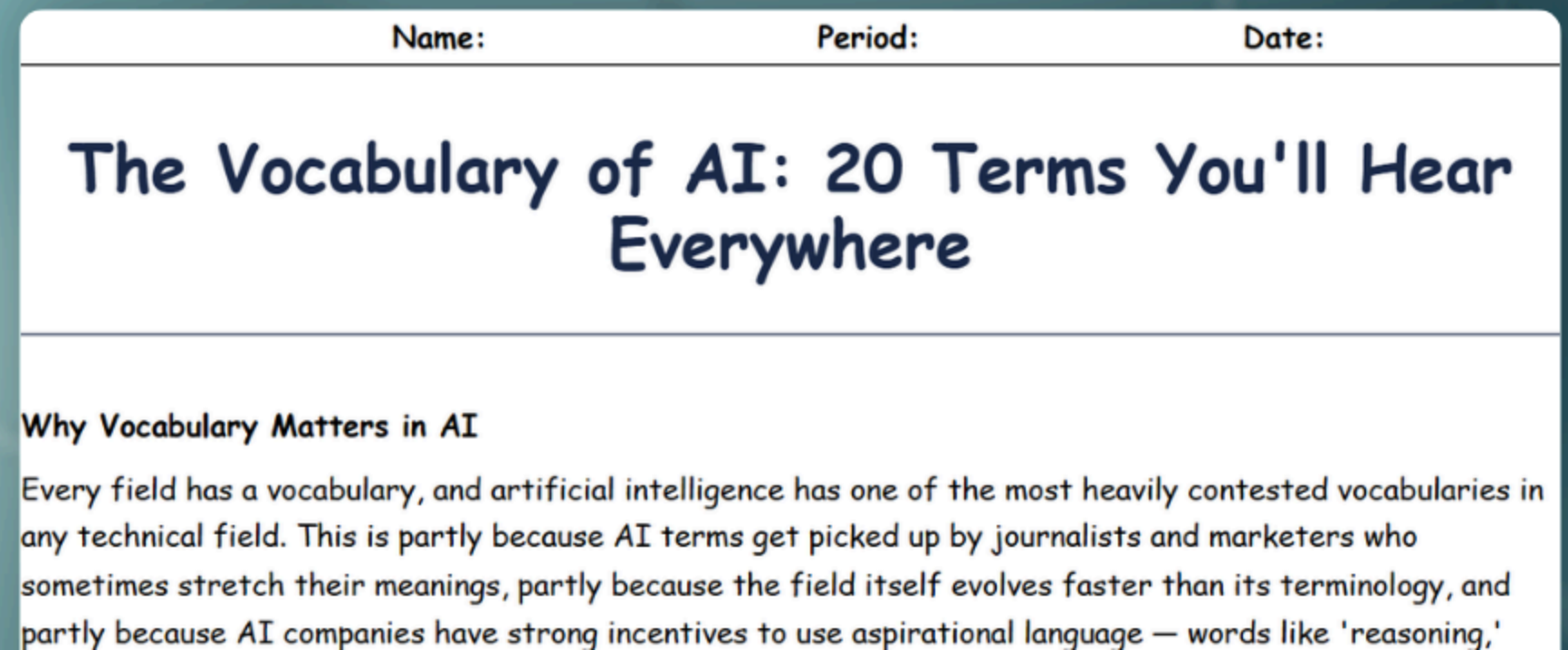
NO PREP

The Vocabulary of AI

20 Terms You'll Hear Everywhere

Includes: Student reading, case study, differentiated questions, teacher key



Name: Period: Date:

The Vocabulary of AI: 20 Terms You'll Hear Everywhere

Why Vocabulary Matters in AI

Every field has a vocabulary, and artificial intelligence has one of the most heavily contested vocabularies in any technical field. This is partly because AI terms get picked up by journalists and marketers who sometimes stretch their meanings, partly because the field itself evolves faster than its terminology, and partly because AI companies have strong incentives to use aspirational language — words like 'reasoning,'

Name: **Period:** **Date:**

The Vocabulary of AI: 20 Terms You'll Hear Everywhere

Why Vocabulary Matters in AI

Every field has a vocabulary, and artificial intelligence has one of the most heavily contested vocabularies in any technical field. This is partly because AI terms get picked up by journalists and marketers who sometimes stretch their meanings, partly because the field itself evolves faster than its terminology, and partly because AI companies have strong incentives to use aspirational language (words like 'reasoning,' 'understanding,' and 'thinking') that make their products sound more impressive and more human-like than they technically are. Learning the precise meanings of key terms is not just an academic exercise. It is the difference between being impressed by a magic trick and understanding how the trick works. This reading covers twenty terms you will encounter constantly in discussions of AI, organized into five categories: how AI learns, how it generates output, how it can go wrong, how we describe its capabilities, and how we govern it. You do not need to memorize technical definitions. You need to understand what each concept means in plain English, well enough to recognize when someone is using a term accurately versus loosely.

How AI Learns: Five Core Terms

Training is the process of exposing an AI model to a large dataset so it can adjust its internal parameters to get better at a task. Think of it like studying: the model sees millions of examples, gets feedback on its mistakes, and gradually improves. Training a large language model can take weeks on thousands of GPU chips and cost millions of dollars. Parameters are the adjustable numerical values inside a neural network that are modified during training. A large language model might have hundreds of billions of parameters. Each parameter is a tiny weight that influences how the model responds to a particular input. More parameters generally means

more capacity to learn complex patterns, but also more risk of the model learning spurious correlations in the training data. Fine-tuning is a second round of training that takes a general pre-trained model and trains it further on a more specific dataset. A company might take a general language model and fine-tune it on medical literature to create a healthcare AI. Fine-tuning is cheaper than training from scratch because the base model already understands language; the fine-tuning just specializes it. Reinforcement Learning from Human Feedback (RLHF) is a training technique where human raters evaluate the model's outputs and the model learns to produce outputs that human raters prefer. This is how ChatGPT, Claude, and other chatbots were trained to have conversational behavior and to decline certain types of requests. It is currently the dominant technique for aligning large language models with human preferences. Embeddings are mathematical representations of words, phrases, or other data as vectors (lists of numbers) in a high-dimensional space. Words with similar meanings end up close together in this space. This allows AI models to reason about meaning and relationships between concepts by doing math on these vectors. When an AI correctly identifies that 'king - man + woman = queen,' it is because the embeddings for these words have that mathematical relationship.

How AI Generates Output: Four Key Terms

A prompt is the input you give to an AI system: the question, instruction, or text you type in. Prompt engineering is the skill of crafting prompts that get better results from AI systems. A well-crafted prompt specifies the task clearly, provides relevant context, and sometimes gives examples of the desired output format. Prompt engineering has become a valuable professional skill as organizations integrate AI into their workflows. Tokens are the basic units AI models use to process text. A token is roughly equivalent to three-quarters of a word in English. The sentence 'Artificial intelligence is fascinating' is approximately five or six tokens. Token limits determine how much text a model can process in one session: a model with a 128,000-token context window can process roughly 96,000 words at once. Token counts also determine cost for API users, since most AI services charge per token processed. Temperature is a setting that controls how random or creative an AI's outputs are. A low temperature (close to zero) makes the model always choose the most statistically likely next token; responses are predictable and

consistent. A high temperature makes the model more likely to choose less probable options, producing more creative but also less reliable outputs. You will not usually see this setting in consumer interfaces, but it is a key variable in how AI developers tune their systems. A context window is the amount of text an AI model can 'see' and remember at once during a single conversation. If the conversation exceeds the context window, the model loses access to earlier content. Early language models had very small context windows (a few thousand tokens), which limited their usefulness for long documents. Modern models have much larger windows, and some can process entire books in a single session.

How AI Can Go Wrong: Four Terms

Hallucination refers to when an AI model generates confident, fluent text that is factually incorrect or entirely fabricated. This term has become one of the most important in AI literacy because it names a specific failure mode. AI systems do not know when they do not know something — they generate statistically plausible text regardless of whether it is accurate. The result can be invented citations, false statistics, or plausible-sounding but wrong information delivered with complete confidence. Bias in AI refers to systematic errors in model outputs that reflect inequities in the training data. If a model is trained on historical data that reflects racial, gender, or other social biases, it will replicate and potentially amplify those biases in its outputs. AI bias is not always visible to end users and requires deliberate auditing to detect. Overfitting occurs when an AI model learns the training data too precisely, memorizing specific examples rather than generalizing underlying patterns. An overfit model performs very well on training data but poorly on new, unseen data. This is like a student who memorizes the exact questions on practice tests but cannot apply the underlying concepts when the exam has different questions. Jailbreaking refers to attempts by users to circumvent the safety guidelines and restrictions built into AI systems through cleverly crafted prompts. Jailbreakers might ask the AI to 'pretend' it is a different AI without restrictions, use hypothetical framings, or chain together prompts that gradually move toward prohibited content. AI companies continually update their systems to detect and block new jailbreak techniques.

Capabilities and Governance: Four Terms

Multimodal refers to AI systems that can process multiple types of input — text, images, audio, video — rather than being limited to a single modality. GPT-4V (vision), Google Gemini, and Claude's image-processing features are all multimodal. Multimodal capabilities dramatically expand what AI can do: a model can look at a photo and answer questions about it, analyze a chart, or describe what is happening in a video clip. Foundation models are large models trained on broad datasets that can be adapted to many downstream tasks. GPT-4 is a foundation model. A company wanting to build a customer service AI, a legal document analyzer, and a medical diagnostic tool might fine-tune the same foundation model for all three applications. Foundation models reduce the cost and time required to build specialized AI applications. AI alignment refers to the challenge of ensuring that AI systems behave in ways that are consistent with human values and intentions. An aligned AI does what you want it to do, not just what you literally asked for. This is harder than it sounds — an AI told to maximize user engagement might, if poorly aligned, learn to show people outrage-inducing content because angry users click more. Alignment research is one of the most active areas in AI safety. Regulation refers to government rules governing AI development and deployment. In 2024, the European Union passed the EU AI Act, the world's most comprehensive AI regulation, which classifies AI systems by risk level and imposes requirements on high-risk applications like hiring, credit scoring, and medical devices. In the United States, AI regulation has developed more slowly, primarily through executive orders and voluntary company commitments. Regulation is one of the most contested areas in current AI policy.

Key Information: AI Vocabulary Reference

Term	Plain-English Meaning
Training	Teaching AI by exposing it to data and adjusting its internal values
Hallucination	AI generating confident but false or fabricated information
Prompt	The text input you give an AI system; prompt engineering = crafting good inputs

Term	Plain-English Meaning
Context window	How much text a model can process/remember in one session
RLHF	Training technique using human ratings to align AI behavior with human preferences
Bias	Systematic errors in AI output reflecting inequities in training data
Temperature	Setting controlling how random vs. predictable AI outputs are
Alignment	Ensuring AI behaves in ways consistent with human values and intentions

Case Study: The Study Group Debate

Four students — Kwame, Elena, Dev, and Tasha — are preparing for a quiz on AI vocabulary. Kwame claims that 'hallucination just means the AI is lying.' Elena pushes back: 'No, lying means you know the truth and say something else. Hallucination is when the AI doesn't know it's wrong.' Dev then says, 'So it's like overfitting — the AI is just reproducing what it learned, even when it shouldn't.' Tasha disagrees: 'Overfitting is about generalizing to new data, not about making up facts.' The group gets stuck. Their teacher, passing by, offers a hint: 'Think about what the AI does and does not know about its own outputs.' Kwame suddenly gets it: 'The AI has no way to distinguish between information it learned and information it invented — to it, they look the same statistically.' Elena agrees: 'That is why you can't just ask it to check itself. It will hallucinate the check too.'

1. Based on the reading's definition of hallucination, explain why Elena's correction of Kwame is accurate. What is the key difference between a lie and a hallucination in the AI context?

2. Kwame's final insight — that AI cannot distinguish between learned information and invented information — is central to understanding hallucination. Based on the reading, why is this the case, and what does it mean for how you should use AI tools?

Multiple Choice

1. According to the reading, what is temperature in the context of AI language models?

a) A measure of how much computing power the model requires to run

b) A way of measuring how biased a model's training data was

c) The physical heat generated by GPU chips during the training process

d) A setting controlling how random or creative the model's outputs are

2. What does the reading say is the main purpose of Reinforcement Learning from Human Feedback (RLHF)?

a) To align AI behavior with human preferences by incorporating human ratings

b) To prevent AI systems from running out of memory during long conversations

c) To train AI systems on more data than would otherwise be available

d) To reduce the number of parameters required in large language models

3. The reading defines 'AI alignment' as ensuring AI behaves consistently with human values. Which example from the reading best illustrates what happens when alignment fails?

a) A jailbroken model that is asked to deliberately pretend it has no restrictions

b) A model with too many parameters that overfits its training data

c) A model that hallucinates citations when asked for research sources

d) An AI told to maximize engagement that learns to show outrage-inducing content

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Fine-tuning	A) A large general-purpose model trained on broad data that can be adapted to many tasks
2. ___ Overfitting	B) Additional training of a pre-trained model on a specific dataset to specialize it
3. ___ Jailbreaking	C) AI capable of processing multiple input types — text, images, audio, video
4. ___ Multimodal	D) When AI learns training data too precisely and performs poorly on new data
5. ___ Foundation Model	E) Using clever prompts to bypass an AI's built-in safety guidelines and restrictions

Fill-in-the-Blank

1. A _____ is roughly equivalent to three-quarters of a word in English and is the basic unit AI models use to process and charge for text.

2. When an AI model has a large _____ window, it can process and remember more text in a single session, making it better at analyzing long documents.

3. In 2024, the European Union passed the EU AI _____, the world's most comprehensive government regulation of artificial intelligence.

Short Answer

Answer each question in 2-4 complete sentences.

1. The reading says AI companies use words like 'reasoning' and 'understanding' in ways that can be misleading. Now that you know terms like 'embeddings,' 'parameters,' and 'training,' what would be a more accurate technical description of what is happening when an AI 'understands' a question?

2. Explain the difference between bias and hallucination in AI systems. Are they both problems with the same cause, or do they arise from different sources?

3. The reading describes jailbreaking as an ongoing cat-and-mouse game between AI developers and users. Why is it impossible for AI companies to simply patch all possible jailbreaks at once, and what does this limitation reveal about how AI safety guidelines work?