FREE RESOURCE · AIEWORKS

Top 50 AI System Design Questions

The exact questions you'll face in ML system design interviews at Google, Meta, Amazon, Apple, Netflix, OpenAI, and Anthropic. No answers — that's in the paid workbook.

50	9	8	3
QUESTIONS	COMPANIES	CATEGORIES	DIFFICULTY LEVELS

Want **complete worked solutions** for all 50 questions? Each answer includes the full framework, step-by-step walkthrough, and what earns points with interviewers.

Get the ML System Design Workbook →

How to use this list: ML system design interviews are 45–60 minutes. You'll be given a vague product problem and asked to design the ML system end-to-end. These 50 questions represent the most common scenarios across FAANG and top AI companies. Study the *categories and patterns* — not just individual questions. The full workbook includes complete solutions with framework, trade-offs, and interviewer rubrics.

Difficulty: Entry / L4 Mid-level / L5 Senior+ / L6+

1. Recommendation & Personalization Systems

Asked at: Netflix, Meta, YouTube (Google), Amazon, Spotify, LinkedIn

1 **Design YouTube's video recommendation system for the home feed of 2 billion monthly active users.**

L5 Two-Tower Candidate Generation

> Focus on multi-stage pipeline: candidate generation → scoring → re-ranking → diversity.

2 **Design Netflix's "You Might Also Like" content recommendation feature at 250 million subscribers.**

L4 Collaborative Filtering Cold Start

3 **Design Spotify's song recommendation system that accounts for real-time listening context (time of day, current mood, recent skips).**

L5 Real-time Features Contextual Bandits

4 **Design Amazon's "Frequently Bought Together" product recommendation for 300 million products.**

L4 Complementary Items Basket Analysis

5 **Design a friend recommendation system for Meta that handles a social graph with 3 billion nodes.**

L6+ Graph ML GNN

> Interviewer wants to see how you handle scale. Graph sampling is key.

6 **Design a news feed ranking system for LinkedIn with 900 million professionals. The system must balance engagement with professional relevance.**

L6+ Multi-objective Ranking Calibration

2. Search & Ranking

Asked at: Google, Amazon, Meta (Marketplace), Airbnb, LinkedIn

7 **Design Google Search's ranking system for 100 billion web pages with 8.5 billion daily queries.**

L6+ Learning to Rank Query Understanding

> Scope to a specific part: query understanding OR document ranking OR freshness — don't try to design all of it.

8 **Design a semantic search system for an internal enterprise knowledge base with 10 million documents.**

L4 RAG Vector Search

9 **Design Airbnb's search ranking system that balances host supply with guest demand across 220 countries.**

L6+ Two-sided Marketplace Personalization

10 **Design an e-commerce product search system with query understanding, typo correction, and faceted filtering.**

L5 NLU Hybrid Search

11 **Design a query auto-complete system that handles 1 billion queries per day with P99 latency under 50ms.**

L5 Low Latency Prefix Matching

3. Ads & Monetization Systems

Asked at: Meta, Google, Amazon, Twitter/X, Snapchat

12 **Design Meta's ad click-through rate (CTR) prediction system for 10 billion daily ad impressions.**

L6+ CTR Prediction DLRM

> Must address feature interaction, sparse embeddings, and online learning.

13 **Design an ad bidding system that optimizes for advertiser ROI while maximizing platform revenue.**

L6+ Auction Theory Bid Optimization

14 **Design a conversion rate optimization system for e-commerce ads that predicts probability of purchase (not just click).**

L5 Post-click CVR Attribution

15 **Design a budget pacing system that ensures advertisers spend their budget evenly over a campaign period.**

L5 Control Systems Feedback Loops

4. Content Moderation & Trust & Safety

Asked at: Meta, YouTube, TikTok, Twitter/X, OpenAI, Anthropic

16 **Design a hate speech detection system for a social media platform with 500 million daily active users across 100 languages.**

L6+ Multilingual NLP Human-in-Loop

> Tradeoff between recall (catching harmful content) vs. precision (over-censoring) is the crux.

17 **Design a CSAM (child sexual abuse material) detection system. What are the accuracy, privacy, and latency requirements?**

L6+ Perceptual Hash Zero False Negative

18 **Design a misinformation detection system for viral news articles shared on a social platform.**

L5 Claim Detection Fact Verification

19 **Design a spam detection system for email with less than 0.1% false positive rate (legitimate email marked as spam).**

L4 Precision-Recall Tradeoff Adversarial

20 **Design a system to detect AI-generated content (deepfakes, LLM-written text) on a content platform.**

L6+ Watermarking Distribution Shift

5. Fraud Detection & Risk Systems

Asked at: Stripe, PayPal, Amazon, Uber, Airbnb, banks

21 **Design a real-time credit card fraud detection system that must make a decision in under 300ms.**

L6+ Real-time Inference Class Imbalance

> 99.9% of transactions are legitimate. How do you handle extreme imbalance and concept drift?

22 **Design a bot detection system for a social media platform to identify automated accounts.**

L5 Behavioral Features Graph Anomaly

23 **Design a seller trust score system for a marketplace that predicts probability of fraudulent listings.**

L5 Entity Scoring Temporal Features

24 **Design an anomaly detection system for financial transactions that can detect novel fraud patterns not seen in training data.**

L6+ Unsupervised AD Autoencoder

6. NLP & LLM-Powered Systems

Asked at: OpenAI, Anthropic, Google, Meta, Microsoft, all tech companies in 2024–2025

25 **Design a RAG (Retrieval-Augmented Generation) system for a customer support chatbot with 10 million support tickets as the knowledge base.**

L5 RAG Architecture Chunking Strategy

> Chunking strategy and reranking pipeline are the discriminating factors here.

26 **Design a system for fine-tuning a large language model on company-specific data while preventing catastrophic forgetting.**

L6+ Fine-tuning Continual Learning

27 **Design a code review assistant that automatically reviews pull requests and suggests improvements, supporting 10+ programming languages.**

L5 Code LLM Tool Use

28 **Design an LLM evaluation pipeline that can automatically measure model quality across safety, factuality, instruction-following, and coding.**

L6+ LLM Eval Red-teaming

29 **Design a document summarization system that handles documents up to 500,000 tokens with high factual accuracy.**

L5 Long Context Hierarchical Summary

30 **Design an AI agent system that can autonomously complete multi-step software engineering tasks (e.g., fix a GitHub issue end-to-end).**

L6+ Agentic AI Tool Use

7. Computer Vision Systems

Asked at: Apple, Tesla, Meta, Google, NVIDIA, autonomous vehicle companies

31 **Design an image classification system for Instagram that categorizes 100 million photos uploaded per day into 10,000 categories.**

L5 · Large Scale Vision · Hierarchical Labels

32 **Design a visual search system for Pinterest that allows users to search by uploading an image and finding visually similar content.**

L5 · Image Embeddings · ANN Search

33 **Design a face recognition system for a mobile device that works on-device with under 50ms inference time.**

L6+ · On-device ML · Model Compression

> Apple-style question. Focus on quantization, CoreML, and privacy-preserving design.

34 **Design an object detection system for retail stores that tracks product inventory via ceiling cameras in real time.**

L5 · Real-time Object Detection · Edge Deployment

8. MLOps, Production, & Infrastructure

Asked at: all companies for senior/staff+ roles, Google, Meta, Amazon AWS

35 **Design a feature store for a company with 500 ML models that need to share features across training and serving with consistent definitions.**

L6+ · Feature Store · Train-Serve Skew

36 **Design a model monitoring system that detects data drift, concept drift, and model degradation across 200 production models.**

L6+ · Drift Detection · Observability

37 **Design an A/B testing platform for ML models that handles interference between experiments and supports multi-variate testing.**

L6+ · Experimentation · CUPED

38 **Design a distributed training system that can train a 70B parameter LLM efficiently across 1,000 GPUs.**

L6+ Tensor/Pipeline Parallelism Megatron-LM

39 Design a model serving infrastructure that can handle 1 million requests per second with P99 latency under 100ms.

L6+ Serving Infrastructure Batching

40 Design an automated ML pipeline (AutoML) that can train, evaluate, and deploy the best model for a given dataset without human intervention.

L6+ AutoML Neural Architecture Search

Bonus: Cross-domain & Applied Questions (41–50)

Mix of applied ML scenarios, emerging topics, and company-specific deep dives

41 Design Uber's surge pricing ML system that predicts demand and adjusts prices in real time for 9 million drivers.

L6+ Demand Forecasting Price Optimization

42 Design a churn prediction system for a SaaS company with 1 million paying customers and 200+ product signals.

L5 Survival Analysis Causal Inference

43 Design a medical diagnosis assistance system that helps radiologists detect tumors in CT scans with higher accuracy than human-only review.

L6+ Medical AI Uncertainty Quantification

44 Design a personalized push notification system that maximizes user engagement while minimizing notification fatigue.

L5 Contextual Bandits Send-time Optimization

45 Design a forecasting system for supply chain demand prediction across 10 million SKUs with seasonal patterns.

L5 Time Series Hierarchical Forecasting

46 Design a multi-modal search system that handles queries containing both text and images (e.g., "find shoes that look like this photo but cheaper").

L6+ Multi-modal CLIP-style Models

47 **Design a privacy-preserving ML system using federated learning for personalizing a keyboard prediction model across 1 billion mobile devices.**

L6+ Federated Learning Differential Privacy

> Apple-style question focusing on on-device learning with privacy constraints.

48 **Design a reinforcement learning system that optimizes datacenter cooling with physical constraints and safety guarantees.**

L6+ Safe RL Model-based RL

49 **Design a responsible AI system that can audit 500 ML models for bias, fairness, and regulatory compliance (GDPR/EU AI Act).**

L6+ AI Fairness Model Cards

50 **Design a next-generation AI assistant (like ChatGPT) from scratch: training pipeline, safety guardrails, serving infrastructure, and human feedback loop.**

L6+ Full-stack LLM RLHF

> OpenAI / Anthropic flagship question. Shows you understand the complete picture from data curation through deployment.