ML System Design Use Cases

1. Visual Search System

Problem Statement & Scope

- **Goal:** Retrieve *visually* similar images for a given query image (like Pinterest).
- **Scope:** Still images only; no video/text input; no personalization.
- **Constraints:** 100–200B images, sub-second latency; only user click interaction for labels.
- **Assumptions:** Results must be ranked by true visual similarity.

ML Task Framing

- **Task:** Learning-to-rank via representation learning (embedding similarity).
- **Input/Output:** Query image → ranked list of visually similar images.

Data Preparation & Feature Engineering

- Images (with optional metadata); user/image click data; synthetic pairs via augmentation.
- Preprocessing: Resize, normalize, standardize; augment for invariance.

Model Design & Architecture

- CNN/ViT Transformer to produce image embeddings;
- Contrastive loss (triplet, InfoNCE, or cross-entropy);
- ANN search (Faiss/ScaNN) for scalable retrieval.

Labeling & Data Collection

- Clicks as implicit positives; augmentation as self-supervised signal;
- Combine with selective manual review for gold standard set.

Evaluation & Metrics

- **Offline:** nDCG, mAP, Precision@k.
- **Online:** CTR, dwell time, user satisfaction.
- **Eval Dataset:** Stratified image pairs with human similarity labels.

Serving & Production

- Query img → preprocessing → embedding → ANN index → nearest/filtered results.
- Background: Index new image embeddings continuously.

Scalability & System Design

- Sharded ANN index; embedding quantization.
- Caching and hot-path optimization for high-traffic queries.

Extensions/Trade-offs/Advanced

- Smart crop, multi-modal search (image + text), iterative retraining.
- Fairness/representation bias detection, handling label noise.

Typ. Interviewer F/U

- Handling new image types, adding metadata, scaling to 10x data, click noise.

2. Google Street View Blurring System

Problem Statement & Scope

- **Goal:** Blur faces/license plates in street images for privacy.
- **Scope:** Only static street view images; process can be offline.
- **Constraints:** High accuracy, low FN on privacy objects.
- **Assumptions:** Annotated dataset of faces/plates; feedback via user reports.

ML Task Framing

- **Task:** Multi-class object detection (faces/plates).
- **Input/Output:** Image → labeled bounding boxes per privacy object.

Data Prep & Feature Eng.

- Human-annotated images for faces/plates;
- Augmentations: scale, rotate, lighting, occlusion.

Model Design & Architecture

- Two-stage detector (Faster R-CNN) or single stage (YOLOv5/7).
- Anchor box tuning for faces/plates; NMS post-processing.

Labeling/Data Collection

- Use internal annotators plus ongoing user reporting to harvest hard examples.

Evaluation & Metrics

- **Offline:** mAP, AP per class, FN/FP rates for privacy objects.
- **Online:** User-reported incidents of unblurred objects.

Serving & Production

- Offline batch: ingest → detect → blur → serve blurred images.
- Human-in-the-loop for escalated reports.

Scalability & System Design

- Distributed data pipeline (batch processing);
- Hardware acceleration for inference (GPU/TPU).

Extensions/Trade-offs/Advanced

- Fairness in detecting diverse faces/plates.
- New object types (eg. street signs);
- Real-time pipeline for sensitive/emergency cases.

Typ. Interviewer F/U

- Handling new privacy targets, dataset bias, multi-country regulatory changes.

3. YouTube Video Search

Problem Statement & Scope

- **Goal:** Retrieve most relevant videos for a user's text query.
- **Scope:** Cross-modal (text→video); short and long videos; no user personalization.
- **Assumptions:** Text meta (title, desc, tags) available for all videos.

ML Task Framing

- Cross-modal ranking (text embedding ↔ video embedding similarity).

Data Prep & Feature Eng.

- Training: Query–Video click pairs; full video content; text meta.
- Features: Subtitle transcript, thumbnail (for advanced multimodal).

Model Design & Architecture

- Text encoding (BERT/Transformer); video encoding (CNN/LSTM/ViT).
- Similarity via cosine/dot product; contrastive learning (dual-tower retrieval).

Labeling/Data Collection

- User click/watch as implicit, curated "gold" for eval.

Evaluation & Metrics

- **Offline:** MRR, Recall@k, mAP.
- **Online:** CTR, watch time, user satisfaction.

Serving & Production

- Index video/text embeddings; query flow: encode→retrieve→re-rank.

Scalability & System Design

- ANN search; distributed retrieval clusters for large video corpus.

Extensions/Advanced

- Add personalization, hybrid token matching, multi-lingual text support.

Typ. Interviewer F/U

- Handling sparse queries, out-of-vocab, new video cold start.

4. Harmful Content Detection

Problem Statement & Scope

- **Goal:** Filter/remove harmful (eg. hate, violence, nudity) posts from platform.
- **Scope:** Multi-modal (text, image, video); fast detection for uploads/comments.

ML Task Framing

- Multi-label, multi-modal classification.

Data Prep & Feature Eng.

- Mod-labeled/flagged data, user reports;
- Features: text, images, user, submit context.

Model Design & Architecture

- Multi-task architecture (shared backbone, task-specific heads);
- Fusion (early/late) for modality combinations.

Labeling/Data Collection

- Human moderators + automated flags as training data.

Evaluation & Metrics

- **Offline:** Precision/Recall, PR-AUC, confusion matrix.
- **Online:** Rate of false positives/negatives, incident reports.

Serving & Production

- Real-time inference; escalate edge cases for human review.

Scalability & System Design

- Batch + real-time; throttled feedback to avoid gaming.

Extensions/Advanced

- Continual learning for adversarial content;
- Explainable outputs for appeals.

Typ. Interviewer F/U

- Handling new threat types, scaling for evolving tactics, appeals workflow.

5. Video Recommendation System

Problem Statement & Scope

- **Goal:** Maximize per-user video engagement via homepage rankings.
- **Scope:** Implicit feedback; extremely high traffic; top N recommendations per user.

ML Task Framing

- Personalized ranking (recommendation).

Data Prep & Feature Eng.

- User/videp behavior logs; social graph; content features (genre, tags).

Model Design & Architecture

- Two-tower neural net; matrix factorization baseline;
- Wide-and-deep, ensemble with content scores.

Labeling/Data Collection

- Implicit (views, likes); explicit (ratings).

Evaluation & Metrics

- **Offline:** nDCG, Precision@k, MAP.
- **Online:** CTR, time watched per session.

Serving & Production

- Candidate generation (ANN or heuristics) → ranking ML.
- Real-time updating for trending content.

Scalability & System Design

- Model/data sharding; real-time feedback loop.

Extensions/Advanced

- Serendipity/diversity loss, controlling for popularity bias.

Typ. Interviewer F/U

- New user/video cold start, trending spikes, filter bubbles.

6. Event Recommendation System

Problem Statement & Scope

- **Goal:** Suggest local/personalized events (e.g., Meetup/Eventbrite).
- **Scope:** Ranked candidate list; implicit and explicit feedback.

ML Task Framing

- Learning-to-rank (pointwise/pairwise) classifier.

Data Prep & Feature Eng.

- Historical event attendance, location, user interests, social features.

Model Design & Architecture

- Binary classifier (attendance or not); feature cross of user × event × context.

Labeling/Data Collection

- Past attended events as positive, unattended as negative for training.

Evaluation & Metrics

- **Offline:** Precision@k, Recall@k.
- **Online:** RSVPs, click/reg rate.

Serving & Production

- Real-time event/date filtering, re-rank candidates.

Scalability & System Design

- Event cache for upcoming/popular events; geo-partitioning.

Extensions/Advanced

- Social graph enhancement, inverse propensity scoring, contextual bandits.

Typ. Interviewer F/U

- New event cold start, sparse interest profiles, locality scaling.

7. Ad Click Prediction

Problem Statement & Scope

- **Goal:** Predict CTR for individual ads on social platforms.
- **Scope:** All users/ads; real-time predictions for auctions.

ML Task Framing

- Binary classification (clicked/not clicked).

Data Prep & Feature Eng.

- User+ad+context features; hot/cold start detection.

Model Design & Architecture

- Wide & deep neural net; heavy categorical embedding and feature crossing.

Labeling/Data Collection

- User clicks/impressions logs.

Evaluation & Metrics

- **Offline:** AUC-ROC, Log loss.
- **Online:** Effective cost per impression (eCPM), CTR.

Serving & Production

- Real-time scoring in ad auctions; latency < 10ms.

Scalability & System Design

- High concurrency; batch scoring for reporting.

Extensions/Advanced

- Bias/fairness (demographics); differential privacy.

Typ. Interviewer F/U

- Ad cold start, fraud/bot detection, feature drift.

8. Similar Listings on Vacation Rental Platforms

Problem Statement & Scope

- **Goal:** Recommend "similar homes" (eg Airbnb, Vrbo) for a given listing.
- **Scope:** All listings; listings may have rich info or only images.

ML Task Framing

- Learning-to-rank via session-based embeddings.

Data Prep & Feature Eng.

- Session logs (what users view together), images, metadata.

Model Design & Architecture

- Embedding via co-occurrence (word2vec-skipgram style) or hybrid image+meta.

Labeling/Data Collection

- Session views/bookings, time-on-page as implicit signal.

Evaluation & Metrics

- **Offline:** Average rank of booked alternative.
- **Online:** CTR for suggested similar listings.

Serving & Production

- Fast retrieval of embedding neighbors; live ranking.

Scalability & System Design

- Data/embedding refresh for new/updated homes; partition by geo/price.

Extensions/Advanced

- Cold start mitigation via metadata boosting; active learning with feedback loop.

Typ. Interviewer F/U

- New listing onboarding, extreme seasonality, diversity constraints.

9. Personalized News Feed

Problem Statement & Scope

- **Goal:** Surface engaging, timely content to maximize user stickiness.
- **Scope:** All posts for all users; multi-objective (clicks, likes, shares, dwell).

ML Task Framing

- Pointwise learning-to-rank or multi-task prediction.

Data Prep & Feature Eng.

- User–post interactions, time features, user and author profiles.

Model Design & Architecture

- Multi-task DNN; hierarchical attention for long interaction history.

Labeling/Data Collection

- Implicit (view, like, comment), explicit (hide/block) labels.

Evaluation & Metrics

- **Offline:** Engagement AUC, nDCG, Like/Share@k.
- **Online:** Session length, activity per visit.

Serving & Production

- Candidate generation → scoring → immediate rerank; cold cache fallback.

Scalability & System Design

- Daily/hourly model re-training; realtime feedback incorporation.

Extensions/Advanced

- Demographic fairness, toxicity filtering, new user bootstrapping.

Typ. Interviewer F/U

- Spam/abuse detection, topic diversity, churn suppression.

10. People You May Know

Problem Statement & Scope

- **Goal:** Suggest possible new connections (eg. LinkedIn, Facebook).
- **Scope:** Link prediction at massive scale; diverse user profiles and activity levels.

ML Task Framing

- Graph link prediction, ranking.

Data Prep & Feature Eng.

- Social graph extraction, past connections, feature similarity.

Model Design & Architecture

- Node2vec/graph neural networks; candidate retrieval then learning-to-rank.

Labeling/Data Collection

- Historic accepted/ignored invitations.

Evaluation & Metrics

- **Offline:** Precision@k, Recall@k.
- **Online:** Connection rate, viewed profile actions.

Serving & Production

- Precompute high-likelihood pairs, cache for fast suggest.

Scalability & System Design

- Partition users, batch compute candidates, scalable fanout for high-degree users.

Extensions/Advanced

- Anti-harassment checks, diversity in suggestions, mutual connections.

Typ. Interviewer F/U

- Cold start for new users, fake/account detection, cross-platform linkage.

Generative AI System Design Use Cases

1. Gmail Smart Compose

Problem Statement & Scope

- Auto-complete email sentences for productivity in Gmail.
- Generates contextually relevant next-phrase suggestions during typing.
- Constraints: Instant, low-latency response; short snippets, privacy of user data.

ML Task Framing

- Conditional text generation (context → next-words).
- Input: User's typed partial sentence.
- Output: Top-k likely continuations.

Data Preparation

- Massive dataset of (email prefix, next phrase) pairs.
- Data filtered for privacy, spelling, toxicity.
- Cleaned, de-identified, deduplicated; rare/typos filtered.

Model Development

- Transformer (seq2seq, e.g., T5, BERT with decoder), pretrained and fine-tuned on email/typing tasks.
- Small, efficient model for fast inference.
- Bias/quality controlled via training, decoding (top-k, nucleus sampling).

Evaluation

- Offline: Perplexity, accuracy on held-out completions, MRR.
- Online: Human-in-the-loop A/B for subjective "helpfulness".
- Key: Low latency, appropriate suggestions.

Deployment/Monitoring

- Model locally in browser/device for privacy, or fast server-side API.
- Monitor abuse, edge cases, model drift; re-train if writing styles change.

Scaling/Advanced

- Personalization signal, multi-lingual, adaptation to user profile.
- Hybrid rule+ML for privacy/redaction.

Interviewer Follow-up

- Handling sensitive/private info, spelling/grammar mistakes, multilingual extension.

2. Google Translate

Problem Statement & Scope

- Translate text between arbitrary languages; real-time (web/app) and batch.
- Support colloquialisms, grammar, code-mixing.

ML Task Framing

- Sequence-to-sequence text generation for N→M languages.
- Input: Source text (Unicode).
- Output: Target text.

Data Preparation

- Large-scale parallel corpora; synthetic data via back-translation.
- Tokenization, normalization, language-specific preprocessing.

Model Development

- Large multilingual Transformer (e.g., mT5, MarianMT, MASS).
- Shared subword vocab for efficient parameter sharing.
- Domain adaptation for specialized texts.

Evaluation

- Offline: BLEU, TER, human assessment for adequacy, fluency.
- Online: User feedback, error logs.

Deployment/Monitoring

- Caching, batching for latency.

- Feedback loop from user corrections, new slang, emergent dialects.

Scaling/Advanced

- Handling rare/low-resource pairs, code-switching, real-time streaming.

Interviewer Follow-up

- Handling ambiguous words, domain adaptation, “zero-shot” translation.

3. ChatGPT: Personal Assistant Chatbot

Problem Statement & Scope

- Free-form conversational agent for general tasks, answering, productivity.
- Open-domain; must not hallucinate or go off-topic.

ML Task Framing

- Causal language modeling + instruction following (LLM).
- Input: Conversation history.
- Output: Next message.

Data Preparation

- Web and dialogue datasets, filtered for quality, style, safety.
- RLHF data for alignment; user feedback collection.

Model Development

- Transformer LLM (GPT-3.5/4), fine-tuned with instruction-following and RLHF.
- Context window management, memory for multi-turn dialogue.

Evaluation

- Offline: Win-rates vs. gold, synthetic testing, hallucination rate.
- Online: Human preference rating, abuse/harms monitoring.

Deployment/Monitoring

- Response filtering, safety layers, detection of prompts for jailbreaks/misuse.
- Model versioning, rapid update for new knowledge.

Scaling/Advanced

- Tool use (plugins), persona adaptation, multimodal support.

Interviewer Follow-up

- Reducing hallucinations, handling safety/adversarial use, non-English/localization.

4. Image Captioning

Problem Statement & Scope

- Generate descriptive text for arbitrary images (e.g., for accessibility or search).
- Multimodal: must handle diverse, real-world imagery.

ML Task Framing

- Vision-to-text generation.
- Input: Image.
- Output: Concise, relevant caption.

Data Preparation

- Curate (image, caption) pairs (COCO, web), filter for relevance/diversity.
- Augment (rotation, crops), dataset balancing.

Model Development

- Encoder-decoder architecture: ViT/ResNet encoder, Transformer decoder.
- Cross-attention; optionally pre-trained on VLM datasets.

Evaluation

- Offline: BLEU, CIDEr, METEOR, SPICE; human rating.
- Online: User preferences, downstream click/engagement.

Deployment/Monitoring

- Latency optimization, batch processing for bulk annotation.

Scaling/Advanced

- Fine-grain region captioning, multi-caption output, low-resource domain adaptation.

Interviewer Follow-up

- Hallucination minimization, regional language support, handling “empty”/ambiguous images.

5. Retrieval-Augmented Generation (RAG)

Problem Statement & Scope

- Enhance generative models with retrieval for up-to-date, factual outputs (e.g., enterprise QA).
- Need reliable, source-grounded responses.

ML Task Framing

- Retrieve-relevant-docs → generative synthesis.
- Input: Query (text), retrieval set.
- Output: Synthesized answer, with citations/sources.

Data Preparation

- Index: Structured/unstructured docs, chunking strategy.
- Retrieval: Dense/sparse embedding indexes.

Model Development

- Dual-encoder: retriever (bi-encoder), generator (LLM with RAG head).
- Train on (query, supporting doc, answer) triples.

Evaluation

- Offline: Retrieval precision/recall, answer faithfulness, groundedness.
- Online: User trust, citation click-throughs.

Deployment/Monitoring

- Real time index refresh, fallbacks for missing info.
- Monitoring for outdated/incorrect info.

Scaling/Advanced

- Complex citations, streaming retrieval, multi-modal retrieval.

Interviewer Follow-up

- Handling contradictory sources, citation attribution, attack resistance.

6. Realistic Face Generation

Problem Statement & Scope

- Generate plausible human faces for content creation, game avatars, etc.
- High realism, concern for misuse (deepfakes).

ML Task Framing

- Unconditional image generation.
- Input: Random (or conditional style).
- Output: Face image.

Data Preparation

- Large, diverse datasets of labeled faces (FFHQ, CelebA), privacy filtering.
- Augmentation: pose, lighting.

Model Development

- GANs (StyleGAN2/3), Diffusion models (SDXL), fine-tuned for realism.
- Latent manipulation for style/morph; quality controls for artifacts.

Evaluation

- Offline: FID, Inception Score, human Turing test (visual realism).
- Online: User rating, misuse detection.

Deployment/Monitoring

- On-demand or batch, monitor for inappropriate content, watermarking for provenance.

Scaling/Advanced

- Editing/editable faces, conditional attributes (age, gender), prevention of misuse.

Interviewer Follow-up

- Ethics/deepfake prevention, generalization to OOD faces, fine-grained controls.

7. High-Resolution Image Synthesis

Problem Statement & Scope

- Generate photorealistic images, e.g., for creative industry, marketing, entertainment.
- Desire for large output (e.g. 4k+, print-ready); must balance quality/clipping.

ML Task Framing

- Unconditional or text/image-conditional generation.

- Input: (Optionally) class label, text, or low-res seed.
- Output: High-res image.

Data Preparation

- High-res, domain-diverse images, augmented across styles/resolutions.
- Upscaling and patch-based splits for training.

Model Development

- Diffusion models (e.g., Stable Diffusion variants, DiT).
- Progressive or two-stage generation (coarse → fine).

Evaluation

- FID, Perceptual scores, expert visual review.

Deployment/Monitoring

- GPU-optimized, pipelined upsampling.

Scaling/Advanced

- Fast upsampling, hybrid local/global attention, style transfer.

Interviewer Follow-up

- Memory/equipment optimization, hallucination in fine detail, new domain adaptation.

8. Text-to-Image Generation

Problem Statement & Scope

- Turn user text prompt into a high-quality, relevant image; broad domain generalization.
- Human-AI collaboration (art, design, idea boards).

ML Task Framing

- Conditional image generation (text2img).
- Input: Prompt.
- Output: Generated image.

Data Preparation

- Massive (image, caption/prompt) pairs (LAION, web), preprocessed/filtered.

Model Development

- Diffusion (Stable Diffusion, DALL-E), CLIP as text-image embedding backbone.
- Prompt conditioning, classifier-free guidance.

Evaluation

- CLIP score, human rating of prompt/image alignment, diversity.

Deployment/Monitoring

- Prompt safety filtering, real-time feedback, style controls.

Scaling/Advanced

- Multi-modal, animated output, fine prompt steering, personalization.

Interviewer Follow-up

- Safeguards, style mixing, rare prompt generalization.

9. Personalized Headshot Generation

Problem Statement & Scope

- Generate realistic headshots for a specific user style/identity (e.g., avatars, professional photos).
- Must match input images; privacy considerations.

ML Task Framing

- Few-shot image generation/identity-preserving synthesis.
- Input: Small set of user photos/prompt.
- Output: New headshots.

Data Preparation

- User-provided samples + augmentation (pose, lighting, background).
- Balanced general set for diffusion personalization (LoRA, DreamBooth, etc).

Model Development

- Fine-tuned diffusion (DreamBooth/Lora), facial ID preservation.
- Very strong regularization to preserve identity, while allowing style transfer.

Evaluation

- Similarity metrics (cosine, CLIP), human review for match and artifact.
- User approval system.

Deployment/Monitoring

- One-off fine-tune per user, privacy and prompt check.
- Retain/discard user data per consent.

Scaling/Advanced

- Style transfer, multi-output sampling for user selection, batch processing.

Interviewer Follow-up

- Preventing “leakage”/copying of likeness, misuse prevention, bias handling.

10. Text-to-Video Generation

Problem Statement & Scope

- Generate short videos from text prompts (e.g. creative, explainer, animation).
- Consistency across frames, content safety.

ML Task Framing

- Conditional video generation, multi-frame synthesis (text2video).
- Input: Prompt, optionally seed/guide images.
- Output: Multi-frame video clip.

Data Preparation

- Large-scale (video, caption) pairs; frame extraction, temporal cropping.
- Curation for safe content, filter for violence/misuse.

Model Development

- Diffusion Transformers or 3D CNNs (Video Diffusion, Sora, Make-A-Video).
- Prompt-to-embeddings mapped to temporal latent space.

Evaluation

- Human rating, CLIP/video alignment, motion quality, frame coherence.

Deployment/Monitoring

- Batch processing, resolution scaling; watermarking for provenance.

Scaling/Advanced

- Prompt steering (action, duration), multimodal/mixed-media generation.

Interviewer Follow-up

- Temporal consistency, safety checks, long-video scaling.