ML Search Interview Cheatsheet

1 Diagnosis Framework (CLIP)

Clarify → **L**ocate → **I**nvestigate → **P**ropose

1. What's the symptom? (latency, accuracy, staleness)
2. Where in pipeline is the bottleneck?
3. WHY is it happening? (root cause)
4. Propose fixes: quick wins → medium → architectural

2 Scenario 1: Latency Bottlenecks

2.1 P50 vs P99 Gap

Large gap = tail query problem (not systemic)

2.2 Model Inference Slow (Tail Queries)

- Feature sparsity → default value computation
- Feature fetch bleeding into inference (waiting)
- Batching irregularity (less common)

2.3 Feature Fetch Slow (Tail Queries)

- Cache misses (tail products not cached)
- Data locality (scattered across shards)
- Missing features → fallback computation

2.4 Fixes (Priority Order)

1. Timeout + degrade to simpler model w/ core features
2. Embed core features in ES (eliminate fetch)
3. Pre-compute ALL features offline
4. Two-stage ranking: cheap model 1000 → expensive model 200
5. Partition by popularity (hot/cold separation)

3 Scenario 2: Offline/Online Metric Gap

3.1 NDCG Good Offline, CTR Bad Online

- Training/serving skew (feature mismatch)
- Metric mismatch (NDCG label ≠ CTR)
- Position bias in training data
- Cold users diluting results (Simpson's Paradox)

3.2 Training/Serving Skew

Fix: Log-and-wait – log features at serving time, use logged features for training

3.3 DCG vs NDCG

Standard NDCG: IDCG from retrieved items only
Problem: [1,0,0,0,0,0] gets NDCG=1.0 (perfect!)
Better: Assume ideal = [1,1,1,1,1,1]

3.4 Position Bias Causes

- Exposure bias (only see top results)
- Click ≠ relevance (users click position 1 by habit)

Fix: Inverse propensity weighting, randomization

3.5 Cold Users (Simpson's Paradox)

Tiered ranking:

- Cold users (<N actions): popularity baseline
- Warm users: hybrid model
- Hot users: full personalized model

4 Scenario 3: Retrieval Recall

4.1 BM25 Missing Relevant Products

- Semantic mismatch (no synonyms)
- Scatter-gather / shard imbalance
- Tokenization issues (hyphens, case)
- No phrase/bigram matching
- Spell correction missing on doc side

4.2 Scatter-Gather Problem

Top-K per shard cutoff → miss relevant items clustered on one shard
Fix: Better shard routing, increase per-shard K

4.3 Hybrid Search (BM25 + Vector)

Combine with RRF:

$$\text{RRF}(d) = \sum_i \frac{1}{k + \text{rank}_i(d)} \quad (k = 60)$$

4.4 When to Use What

- BM25: navigational, exact match ("Nike shoes")
- Vector: semantic, exploratory ("cozy home decor")
- Hybrid: union results for max recall

4.5 Reduce Hybrid Latency

- Route queries (BM25 for navigational)
- Run parallel, not sequential
- Offline semantic → synonym lists → BM25 only
- Early termination if BM25 confident

5 Scenario 4: Model Degradation

5.1 Performance Drops Over Time

- Stale model (not retrained)
- Feature drift (distributions changed)
- New products (no signals)
- New users (sparse history)
- Seasonality

5.2 Feature Drift vs Training/Serving Skew

- Feature drift = symptom (distributions change)
- Training/serving skew = problem (train ≠ serve)
- Log-and-wait = solution

5.3 Cold Start Products (Feedback Loop)

No signals → low rank → no exposure → no signals
Exploration strategies:

- Reserved slots (positions 8-10 for new items)
- Epsilon-greedy (10% random new product)
- Thompson Sampling (uncertainty bonus)
- Time-decayed boost (fades over 7-14 days)

5.4 Retraining Cadence

Weekly or monthly, depending on drift rate
Trigger on: performance drop, major catalog change, goal change

6 Scenario 5: A/B Test & Multi-Objective

6.1 Metric Disagreement

CTR up, GMV down → investigate before shipping!

- Clickbait (high CTR, no conversion)
- Cheap products ranked higher
- Cannibalizing discovery

6.2 Guardrail Metric

Metric that must NOT degrade even if primary improves

Example: CTR is primary, GMV is guardrail

6.3 Multi-Objective Optimization

- Weighted combination: $\alpha \cdot \mathrm{CTR} + \beta \cdot \mathrm{GMV}$
- Revenue-weighted labels: click = GMV value (not 1)
- Multi-task learning (shared representations)

6.4 Revenue-Weighted Labels Downside

Sparse signal: most clicks don't convert → 95% zeros

6.5 Tuning Weights

1. Normalize scores to same scale
2. Start with business intuition (0.5, 0.5)
3. Grid search offline
4. A/B test top candidates online

6.6 Pareto Frontier

Set of solutions where you can't improve one objective without hurting another

Present to product team: "Pick a point based on business priority"

7 Key Formulas

DCG@K:

$$\mathrm{DCG@K} = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

NDCG@K:

$$\mathrm{NDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}$$

RRF:

$$\mathrm{RRF}(d) = \sum_{r \in \mathrm{rankers}} \frac{1}{k + r(d)}$$

Statistical Significance:

$p < 0.05$ = 95% confident result is not random chance

8 Quick Reference

Problem	**First Question to Ask**
High P99 latency	P50 vs P99 gap?
Offline/online gap	Training/serving skew?
Low recall	BM25 or semantic issue?
Model degradation	When last retrained?
Metric disagreement	What's the guardrail?

Term	**Meaning**
Log-and-wait	Log serving features, train on them
Simpson's Paradox	Subgroup trend reverses when combin
Feedback loop	No signal → no exposure → stuck
Pareto frontier	Best tradeoff curve
Guardrail metric	Must not degrade