

EXERCISE IDEA & DESCRIPTION

Work in groups of ~3 people, each working on a single topic. Each one covers a use case that requires ML system design and proper MLOps techniques.

Topics are described with main idea, client, general goal of the system, as well as set of assumptions you can make and example considerations for design.

You can, and are expected to, **ask questions**, clarify things to consider, or assumptions to make. Make assumptions if you need, you don't have to ask about every small detail, but imagine asking the client about major things about the product that they want.

Prepare the following:

- rough sketch of system components and how data / requests flow between them
- technologies (at least roughly) that you would use for particular elements
- assumptions you made, things you considered, motivation for design and technology choices
- short presentation of results, in whatever form you want: slides, Word document, Figma board, whiteboard sketch + physical notes

After some time for preparation, we will present and discuss the results. **There are no absolutely correct answers!** The important thing is to consider tradeoffs, future of the system, motivate choices, and make sure you wouldn't trip over some simple thing when actually building the system.

General considerations:

- how easy is it to build
- maintainability
- cost now & in the future
- scalability
- data availability
- technologies pros & cons (e.g. use lecture slides)
- DevOps, security, monitoring, logging, privacy

Topic 1: AI girlfriend

Idea: Personalized LLM-based partner, commonly known as "AI girlfriend / boyfriend".

Client: Many single people, business-to-customer (B2C).

Goal: Conversations partner, personalized assistant.

Assumptions:

1. User can manually pre-configure options through UI, e.g., history, interests.
2. System should integrate with files, e.g., read self-made poetry from Word docs.
3. Model must have configurable character and tone, keep memory, and act appropriately.
4. Persona can have limited world knowledge, mathematical reasoning, coding etc.

Example considerations:

1. Data privacy - we are dealing with personal, sensitive data.
2. Computational power - LLM inference, real-time inference, model size, memory needs.
3. Memory management - how to store and manage memories?

Topic 2: Product recommender

Idea: Recommendation engine for computer games store, like Steam / Epic Games / GOG.

Client: Store company (product owner), individual people.

Goal: More products sold = more money, users want to discover and engage with novel & interesting games.

Assumptions:

1. You have access to rich metadata about current & upcoming games.
2. Large-scale historical data, sales, trends etc. is available to you.
3. You have rich user personalization information, e.g., games library, play times, reviews.

Considerations:

1. Scalability - hundreds of millions of users, tens of thousands of games.
2. Timing - when to retrain models? When to make predictions?
3. Business rules - how to include paid promotions, promote new large releases etc. that bring additional revenue?

Topic 3: IoT predictive maintenance

Idea: Predict possible machine failure to stop them and perform maintenance

Client: Industrial client with computerized machinery, including IoT sensors

Goal: Predict if each device will probably encounter a mechanical failure in next week

Assumptions:

1. You have a lot of sensors, very frequent data gathering, and many features for each machine (e.g., physical, mechanical, electrical measurements).
2. Some sensors have more historical data than others.
3. Constant monitoring is not necessary, only weekly recommendations.

Considerations:

1. Class definition - how to define target metric and perform labeling?
2. Scheduling - how often to retrain and make predictions? How to schedule this?
3. Scalability - how to scale the system?

Topic 4: Credit scoring

Idea: Algorithmically estimate credit risk, give credit or not.

Client: Banks, individual clients.

Goal: Maximize number of people that get credit, but also minimize risk to bank.

Assumptions:

1. We have rich financial history, including internal & external sources, such as debtors registries.
2. Significant amount of raw data available, with multi-year history.
3. External legal factors & regulations, such as consumer rights, credit policy (e.g. legally required buffers).

Considerations:

1. Explainability - client is entitled to explanation if we reject him, how can we do that?
2. Data privacy - how do you process and store data, what kinds?
3. Bias - mitigating sensitive variables usage like gender or age.

Topic 5: Intelligent tour guide

Idea: Mobile app to provide contextual audio guide based on e.g. chat questions, location, pictures of things.

Client: Individual tourists.

Goal: Enhance tourism experience, suggest places to visit and buy tickets to.

Assumptions:

1. We have access to a lot of data to enhance capabilities, e.g. previous & current locations, external integrations such as tourist services, Wikipedia, image search.
2. Access to preferences, prior visits, country of origin, previous reviews, etc.
3. We can access or integrate external ticketing systems.

Considerations:

1. Computational power - if you use heavy models like LLMs or TTS/STT, where do you compute?
2. Hallucinations and safety - how do you ensure exact, relevant guides?
3. Evaluation - how do you measure success of your guide system?