**Team Two**

*ACME company is a growing community of online video creators. Currently user retention is very low however for high retention users, the time spent on the platform is extremely high. The product strategy team are confident that if users were served content that is more relevant and is aligned to their interests, then it would increase user retention across the board. ACME company need your help creating a content recommendation system to allow users to find relevant video content creating a more tailored experience for end users.*

**Key Stakeholders**

**VP of Engineering**

The end user of this system would be users of the video platform. We would like a recommendation system using a new technology we’ve been hearing about called Generative AI. Using this system, we would like real time recommendations for users based on their prior viewing history and the interests they have selected when signing up to the platform. It is critical that inferences are loaded in a very performant manner, we can’t tolerate a response time of more than 500 milliseconds from the model’s inference endpoint. Furthermore, this system has to be up 24x7 with no downtime.

Currently, the platform has 400,000 users. Currently, low retention users will spend 10 minutes on the site during which time, they will click around on about 30 different pages, on each page, we expect recommendations to be loaded. Typically, in a given month, we have about 100,000 low retention users.

**Head of Data Science**

We have a data science and engineering team here at ACME with 30-35 members. They are very experienced at developing ML models and have a background in the financial services industry, entertainment and AdTech. They are comfortable with Python and Java programming languages and we would like to understand how this model would be deployed using AWS.

**Chief Financial Officer**

We realise the importance of a good recommendation system for our platform and appreciate that it is expensive to operate, from a financial perspective, we are willing to allow production systems to have very powerful infrastructure however, we are very keen that strict budgets are enforced and monitored in non-production environments.