LIG106

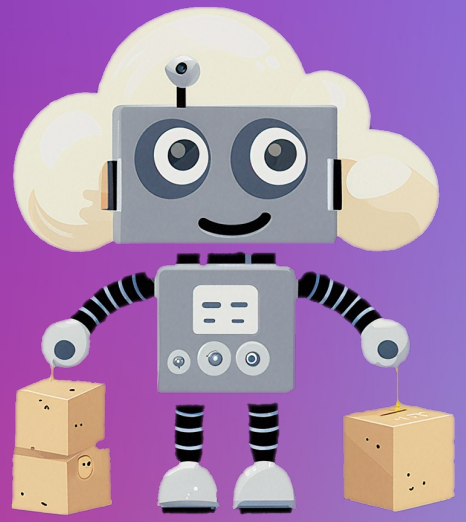
An Executive Playbook for Choosing the Optimal Model

Juan Sanz

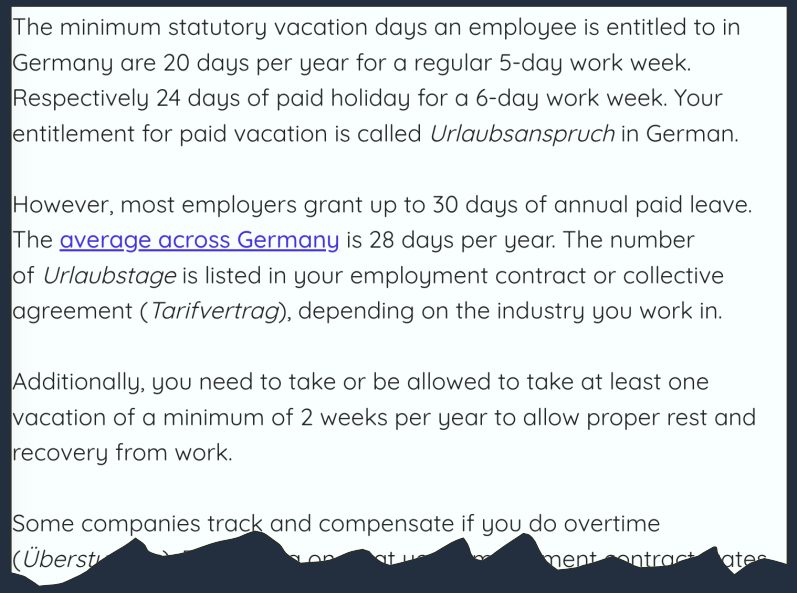
Solution Architect
AWS

Roger Weber

Solution Architect
AWS



Define Scenario – Model leaderboards are a great starting point but may not apply to your use case


The minimum statutory vacation days an employee is entitled to in Germany are 20 days per year for a regular 5-day work week. Respectively 24 days of paid holiday for a 6-day work week. Your entitlement for paid vacation is called *Urlaubsanspruch* in German.

However, most employers grant up to 30 days of annual paid leave. The average across Germany is 28 days per year. The number of *Urlaubstage* is listed in your employment contract or collective agreement (*Tarifvertrag*), depending on the industry you work in.

Additionally, you need to take or be allowed to take at least one vacation of a minimum of 2 weeks per year to allow proper rest and recovery from work.

Some companies track and compensate if you do overtime

User: How many PTO days do you have in Germany?

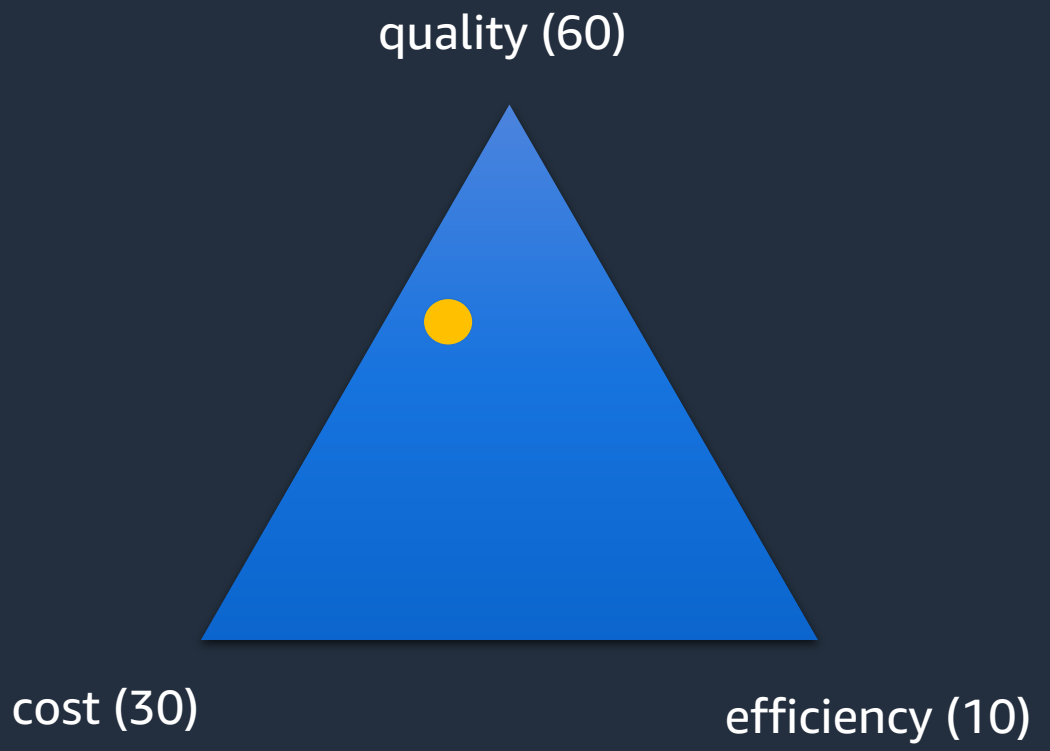
- Language
- References to sources
- Sensitive data, compliance
- Affordable solution
- 2000 users (target population)
- Minimal engineering effort
- Response times (interactive)

Define model constraints and evaluation objectives

Model Constraints:

- ✓ Multilingual support (English, German)
- ✓ Secure and responsible AI, regional availability
- ✓ Managed service with shared model
- ✓ 20k questions expected per day (2k input tokens => 100 output per query)
- ✓ Less than $500/day

Evaluation Objectives & Weights



Service model selection

	Name	Max Tokens	Price per 1,000 input tokens	Price per 1,000 output tokens	Languages
✓	Jurassic Ultra	8,000	$0.0188	$0.0188	en, es, fr, de, pt, it
	Cohere Command	4,000	$0.0015	$0.0020	en
	Cohere-Light	4,000	$0.0003	$0.0006	en
✓	Claude Instant	100,000	$0.00080	$0.00240	en (+others)
✓	Claude v2.1	200,000	$0.00800	$0.02400	en (+others)
	Llama2 (13B)	4,000	$0.00075	$0.00100	en
	Llama2 (70B)	4,000	$0.00195	$0.00256	en
✓	Titan Express	8,000	$0.0003	$0.0004	en (+others)
	Titan Lite	4,000	$0.0008	$0.0016	en

Model Constraints:

- ✓ Consume from Amazon Bedrock
- ✓ Minimum 8k tokens required
- ✓ English & German supported

Define how to measure quality, cost, and efficiency

Claude v2.1

In Germany, you are entitled by law to a minimum of 20 paid vacation days per year if you work a 5-day week. You can also use public holidays strategically with vacation days to maximize your time off.

Titan Express

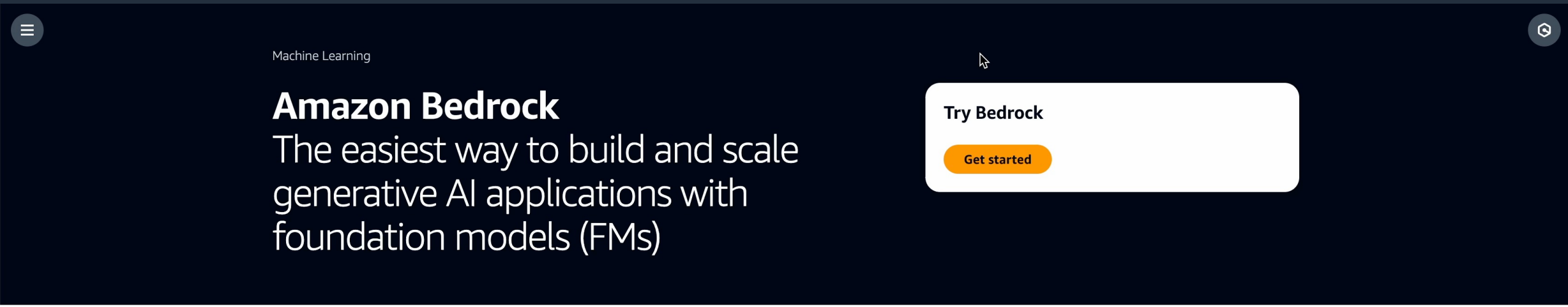
In Germany, employees are entitled to a minimum of 20 paid vacation days per year.

Pairwise model evaluation with Bedrock

Model	Quality*	Cost	Efficiency
Claude v2.1	1132 (+9)	$920	< 2s
Claude Instant	1109	$92	< 0.5s
Titan Express	1098 (-9)	$88	< 0.5s
Jurassic Ultra	1203	$1974	< 2s

* illustrative ELO ratings, not based on a real experiment

Model evaluation



Overview

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

Benefits

Accelerate development of generative AI applications using FMs through an API, without managing infrastructure.

Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case.

Use AWS tools and capabilities that you are familiar with to deploy scalable, reliable, and secure generative AI applications

Please complete the session survey.

Thank you!

Juan Sanz
jusanj@amazon.ch

Roger Weber
drweb@amazon.ch