Carnegie Mellon

Overcoming the Challenges of Large-Scale AI Deployment

Len Bass

AI history is full of optimistic predictions

- **1958**
- H. A. Simon and Allen Newell: "within ten years a digital computer will be the world's chess champion"
  - Deep Blue beat Gary Kasparov in 1997

Carnegie Mellon

Optimism about AI systems persists

- **2024**
- 80% of AI projects do not go into production
  - https://www.rand.org/pubs/research_reports/RRA2680-1.html (2024)

- This is **twice** the rate for non AI projects.

Carnegie Mellon

Why do AI systems fail to get into production so frequently?

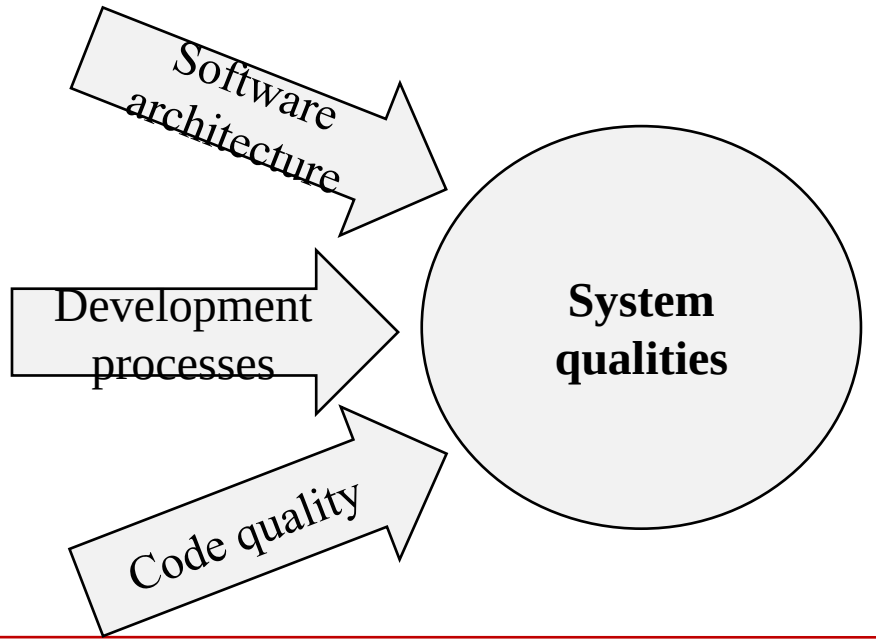
- **Achieving quality in AI systems is difficult**
- More interdisciplinary
- AI systems are based on statistical models

System quality

- Quality is fitness for use
- Characterized by various dimensions – called quality attributes. E.g.
  - Performance
  - Security
  - Reliability
  - …

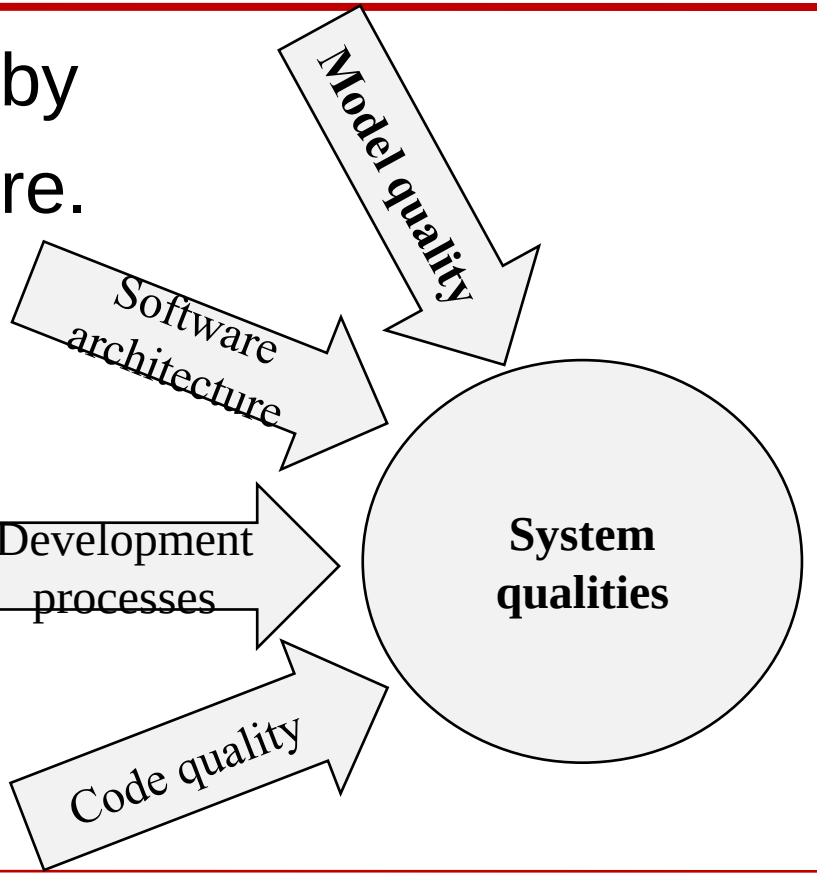
In non-AI systems

- Quality is determined by
  - Software architecture.
  - Development processes
  - Code quality



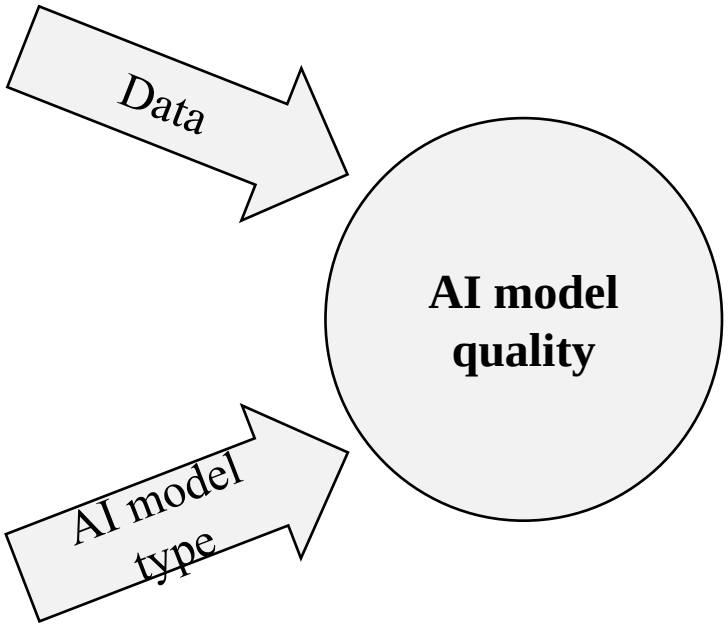
In AI systems

- Quality is determined by
  - Software architecture.
  - Development processes
  - Code quality
  - **Model quality**



Model quality

- Model quality depends on
  - AI model choice
  - Data



Data quality is important

- Sufficient training data – 10x number of parameters in the model
- Distribution of training data
  - Reflect distribution of requests
  - Corrected for biases
- Adjusted to reflect security concerns

Problems with data

- Data drift – training data is not representative of actual data after some period of time. E.g. housing prices have changed
- Environmental change – e.g. training data represents housing prices in urban areas, attempt to use model for rural areas.
- Regulatory changes – ongoing regulations can affect issues such as fairness, transparency, accountability, and human oversight

Mitigating data problems

- Data drift
  - Continual monitoring during operations
  - Retraining the model if necessary
- Environmental drift
  - Same as for data drift – monitoring and retraining as necessary

Mitigating data problems

- Regulatory change
  - Regulatory watch – an organizational unit should monitor ongoing regulation change and notify developers when landscape changes

Mitigating model modifications

- Model code changes
- Models may be changed as a result of data problems
- Use software architecture to isolate model API through a translation layer.

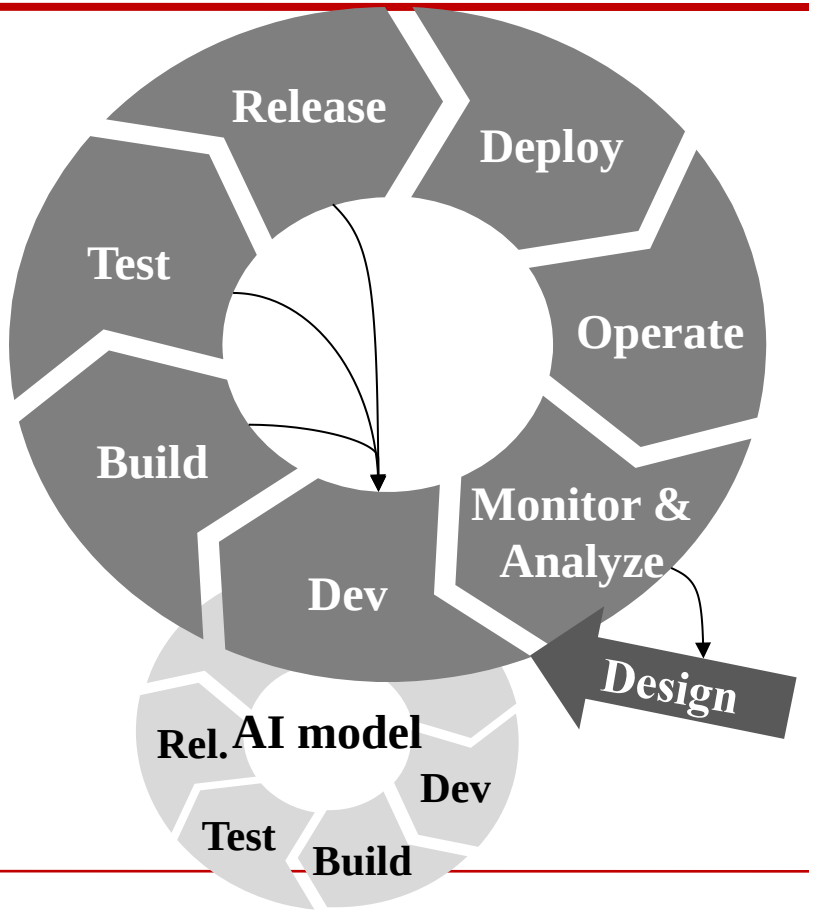
Furthermore

- Definitions of quality attributes are expanded in AI systems
- Performance includes latency, throughput, and accuracy
- Security includes attacks on data – e.g. poisoning the data
- …

Carnegie Mellon

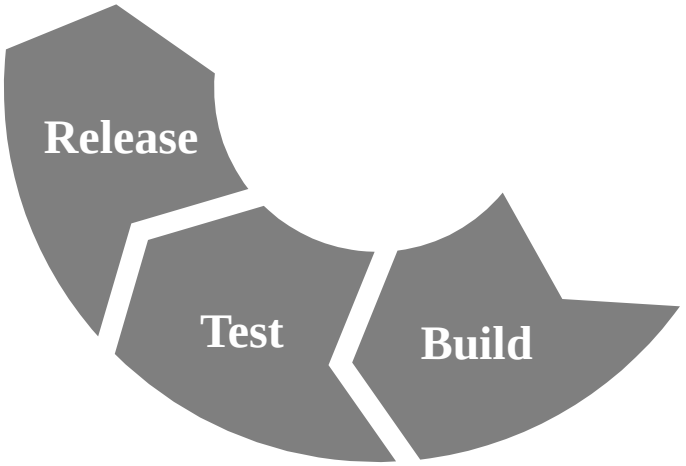
Development practices impact quality

- Note semi-circle at the bottom
- These are additional activities to prepare an AI model



AI model preparation

- Activities include
  - Preparing the data
  - Developing the model
  - Testing and evaluating the model
  - Releasing to build stage



Impacting quality

- Any of the activities in the model preparation can impact quality
  - Data preparation involves data cleaning, resolving missing values, and dealing with outliers
  - Model build involves training and selecting features
  - Test must check for data distribution and bias introduction

Tool support

- Many tools exist to support the model development life cycle
  - Data cleaning
  - Data lineage
  - Model choice and packaging
  - …

Quality in AI systems (summary)

- Achieving quality in AI systems is more difficult than in non-AI systems because of the addition of data quality issues.
- These are manifested in
  - Quality attributes achievement
  - Additional practices to prepare the model
- Quality is achieved by developers mitigating against potential problems.

Why do AI systems fail to get into production so frequently?

- Achieving quality in AI systems is difficult
- **More interdisciplinary**
- AI systems are based on statistical models

Expertises needed

- Building large scale AI systems requires a variety of expertises:
  - Software engineering
    - Application domain
    - Infrastrcture and tooling
  - Data science and engineering
    - Building models and processing data
    - Infrastructure tool ing for models and data

Interdisciplinary teams

- Development teams are inherently interdisciplinary because of the variety of expertises required

Problems with interdisciplinary teams

- Communication barriers: People from different fields may use different terminology or have varying communication styles, leading to misunderstandings.
- Cultural clashes: Different disciplines may have distinct cultural norms or values, which can create friction within the team.

Problems with interdisciplinary teams

- Power struggles: Members of dominant disciplines may assert control, leading to resentment or a lack of collaboration.
- Resistance to change: Members of established disciplines may be resistant to new ideas or approaches from other fields.

Mitigating interdisciplinary problems

- Education and training in unfamiliar disciplines
  - Learn vocabulary and concepts
  - Understand culture
- Time to build teams
  - Time for education and training
  - Time to make teams effective

Tuckerman's model of team development

- "Coming together is a beginning. Keeping together is progress. Working together is success."
- Characterized as forming, storming, norming, and performing

Why do AI systems fail to get into production so frequently?

- Achieving quality in AI systems is difficult
- More interdisciplinary
- **AI systems are based on statistical models**

Different type of models

- Narrow ML models designed to support particular tasks
- Foundation models with no specific task designation

Narrow ML models

- What are they?
- What are they used for?
- What are the problems with narrow models?
- What are the mitigations?

What are Narrow ML models?

- Statistical models
- Intended for a specific task
- Input to a narrow ML model is a data item consisting of a set values of labelled independent variables.
- Output is a generated value of a dependent value.

Spam filter example

- Input is
  - email message,
  - sender's email,
  - recipients email,
  - subject line,
  - Email headers
  - …
- Ouput is: spam or not spam

What are narrow ML models used for?

- Suitable for three main purposes
  - Classification – assigns a category to an input. e.g. this email Is spam
  - Regession – returns a continuous value to an input. E.g. This particular process will take 3 days to complete
  - Clustering – groups similar items. E.g. this data set has 10 groups, clustered by age.

What are the problems with narrow models?

- Ethical concerns
- Interpretability and explainability
- Generalization and overfitting
- Robustness and adversarial attacks

Mitigating ethical concerns and bias

- Diverse and Inclusive datasets:
- Representation: Ensure that the training data represents a diverse population to avoid perpetuating existing biases.
- Ethical Frameworks
- Ethical Review Boards
- Bias Detection Tools

Mitigating interpretability and explainability

- Explainable AI (XAI) Techniques:
  - LIME (Local Interpretable Model-Agnostic Explanations)
  - SHAP (SHapley Additive exPlanations)
  - Visualizations:
- Feature Importance

Mitigating generalization and overfitting

- Regularization
- Cross-Validation:
- Data Augmentation
- Feature Engineering
- Relevant Features
- Hyperparameter Tuning

Mitigating robustness and adversarial attacks

- Adversarial Training
- Input Validation and Sanitization
- Feature Noise Injection
- Regularization Techniques

Regulatory and legal challenges

- Regulations: The rapid development of AI has outpaced the creation of comprehensive regulations, leading to uncertainty and potential legal issues.
- Liability: Determining liability in cases involving AI-powered systems can be complex, especially when accidents or damages occur.

Foundation Models

- What are they?
- What are they used for?
- What are the problems with foundation? models
- What are the mitigations?

What are Foundation Models?

- A foundation Model (FM) is trained on an extensive and diverse dataset, often comprising billions or even trillions of data points.
- The training data is largely unlabeled,
- FMs are general purpose but can be customized for particular applications.
- Large language models (LLMs) are a type of FMs.

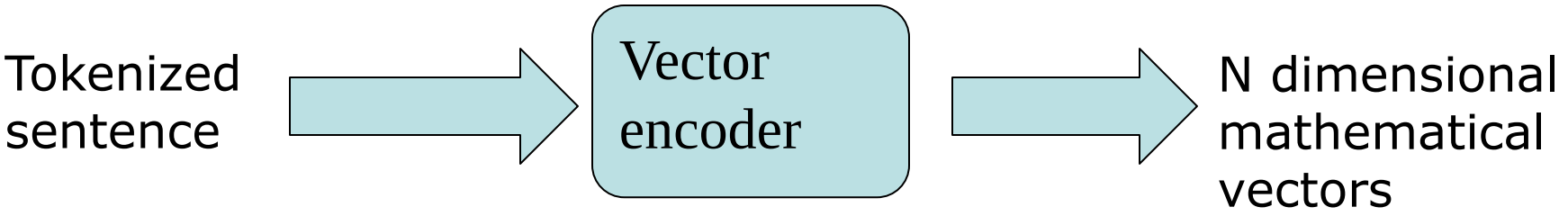
What are Foundation Models used for?

- Natural Language Processing
- Text Machine translation
- Question answering
- Summarization
- Image generation and classification
- Object detection
- Code Generation:
- Other Applications – still being explored.

Transformer architecture

- Primary basis for foundation models
- Two key concepts
  - Vector spaces
  - Attention mechanism.

Vector space



- Sentence is broken into tokens (think syllables)
- Each token is represented as a vector
- Sentence is represented as a collection of vectors
- Dimension of vectors may be in the thousands

Attention mechanism

- Given a sentence, the tokens in the sentence define a collection of regions in vector space.
- Attention is the process of analyzing those regions and their order to extract meaning from the sentence.

Customizing

- Foundation models are trained on unlabeled data. They must be customized for specific tasks.
- They can be customized by adding context information to the input query.

Adding context information

- Add it through editing the prompt. – prompt engineering
- Add it from supplemental data sources – Retrieval Augmented Generation (RAG).

What are some problems with Foundation Models?

- Data Privacy and Security
- Sensitive Data Exposure
- Misuse and Misinformation
- Deepfakes and Fake Content

Guardrails to mitigate problems with FMs

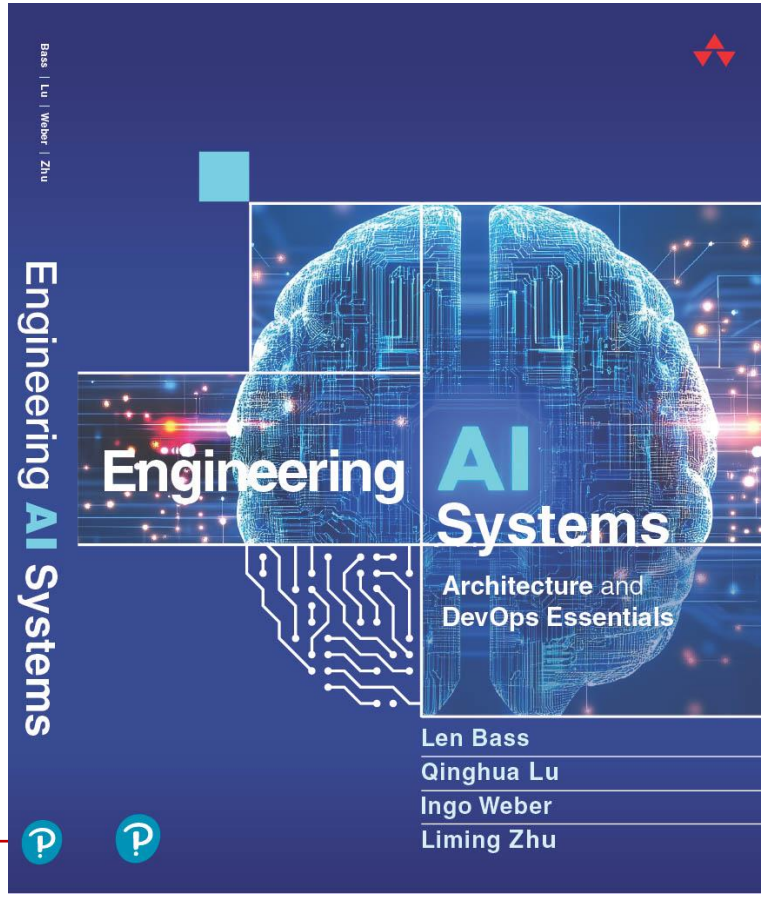
- A guardrail is a component that monitors input or output from an FM
- Can reject input that asks for personal or sensitive information
- Can reject input that attempts to add midinformation
- Can reject output that contains personal, sensitive data or misinformation.

Summary

- Three contributors to failure to get AI systems into production
  - Achieving quality in AI systems is difficult
  - More interdisciplinary
  - AI systems are based on statistical models
- Recognizing problems and mitigating them should improve percentage of AI systems that get into production.

More information

- Engineering AI Systems
  - To be published in early 2025



- **Questions?**