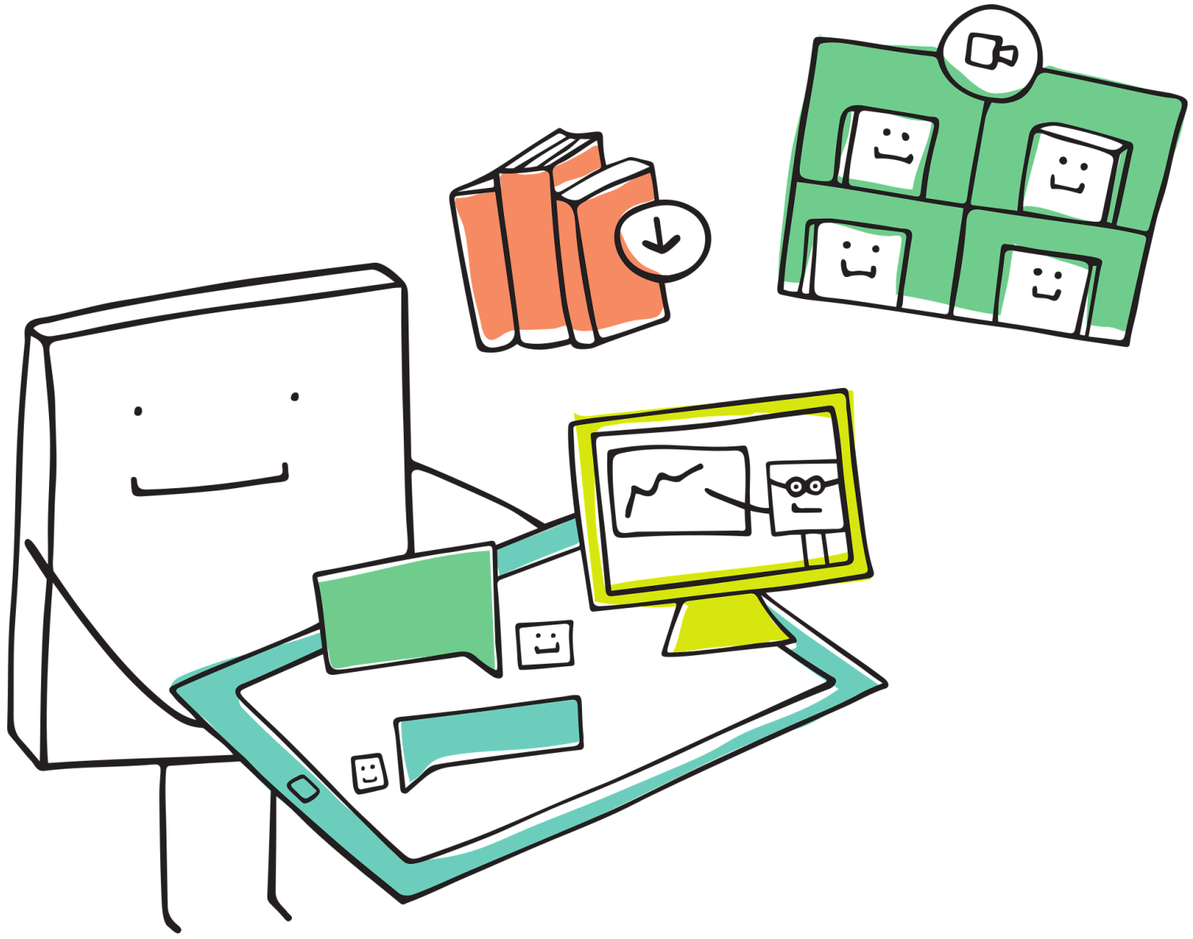
Overview of LLMOps

LLMOPS CONCEPTS

Max Knobbout, PhD
Applied Scientist, Uber

datacamp

What we will learn in this course



- LLMOps helps to effectively **manage**, **deploy**, and **maintain** large language model applications
- We will cover:
  - Fundamentals
  - Lifecycle of LLM applications
  - Challenges and considerations

[1] Illustrations by Manfred Steger @ Pixabay

A recap of LLMs

What are LLMs?

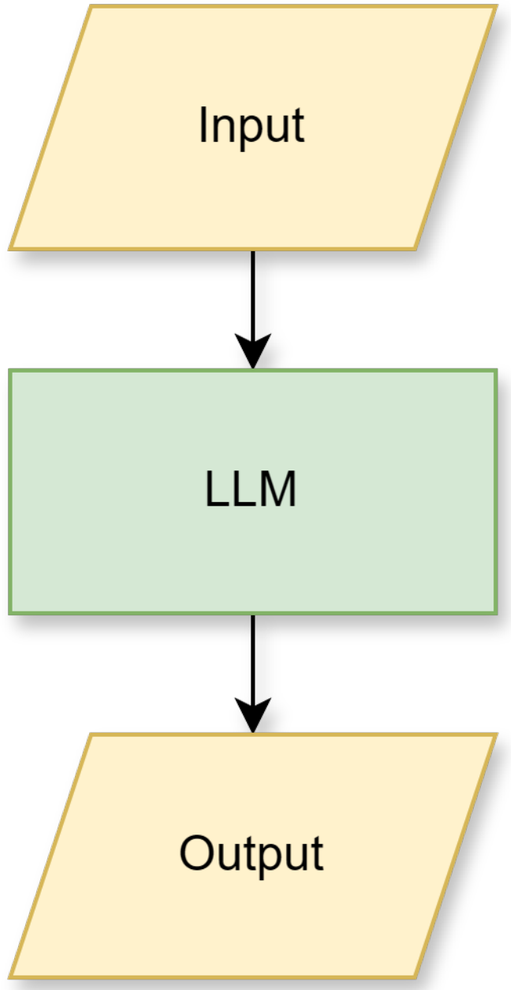
- Trained on extensive text data
- Can understand and generate human-like text
- Represent an AI breakthrough

What sets them apart?

- Typically pre-trained
- Massive number of parameters
- Significant computational resources
- Unpredictable

How it started...

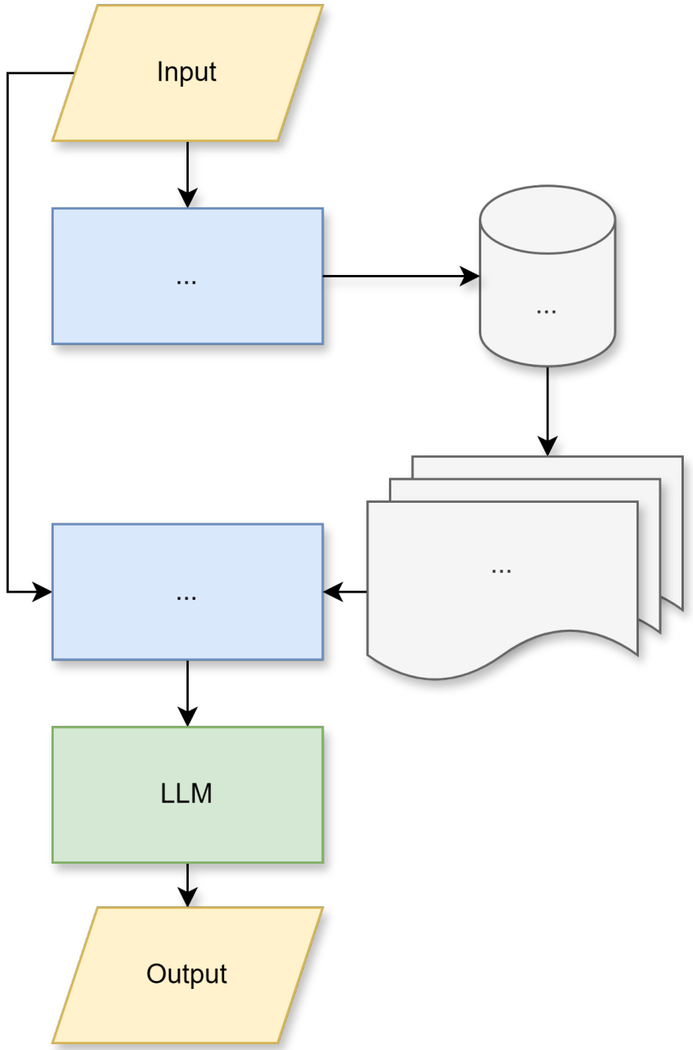
- Queries were directly fed into the model
- The focus was on operating the model
- Only when the model was fine-tuned data was introduced



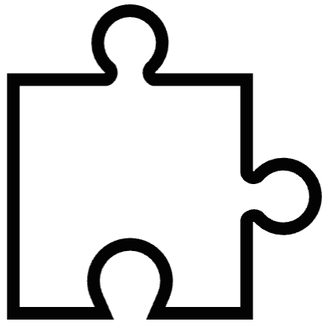
... versus how it's going

- Integrating organizational data before text generation
- Steps can involve data processing and manipulation
- One, or multiple model calls, accommodating text, image, or multi-modal

Resulting in what we call **LLM applications** throughout this course

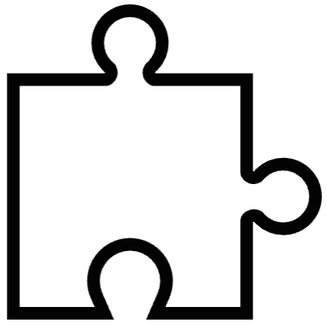


The need for LLMOps

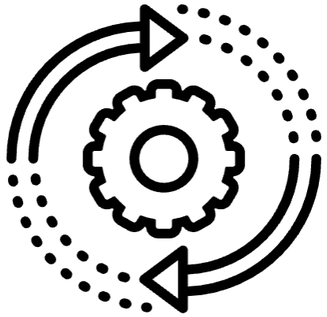


Seamless integration

The need for LLMOps

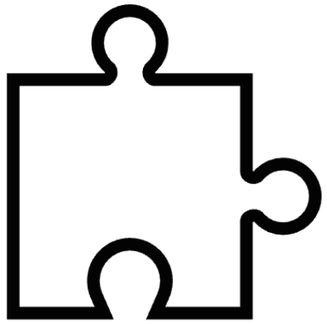


Seamless integration

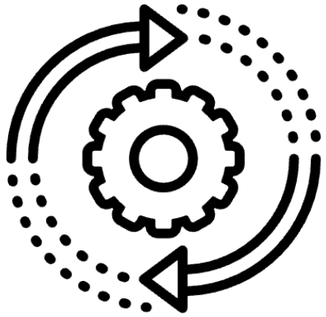


Smooth lifecycle transitions

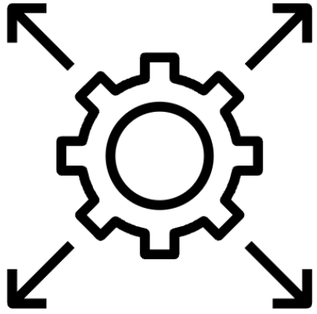
The need for LLMOps



Seamless integration



Smooth lifecycle transitions



Efficient, scalable management

LLMOps versus MLOps

LLMOps versus MLOps

Some differences:

	LLMOps	MLOps
Model size	Large	Typically smaller
Data	Text	Any data
Pre-trained models	Typically yes	Typically no
Model improvement	Prompt engineering & fine-tuning	Feature engineering & model selection
Generalization	General-purpose	Fixed scope
Unpredictability	High	Low
Output	Primarily text	Task-specific

Let's practice!

LLMOPS CONCEPTS

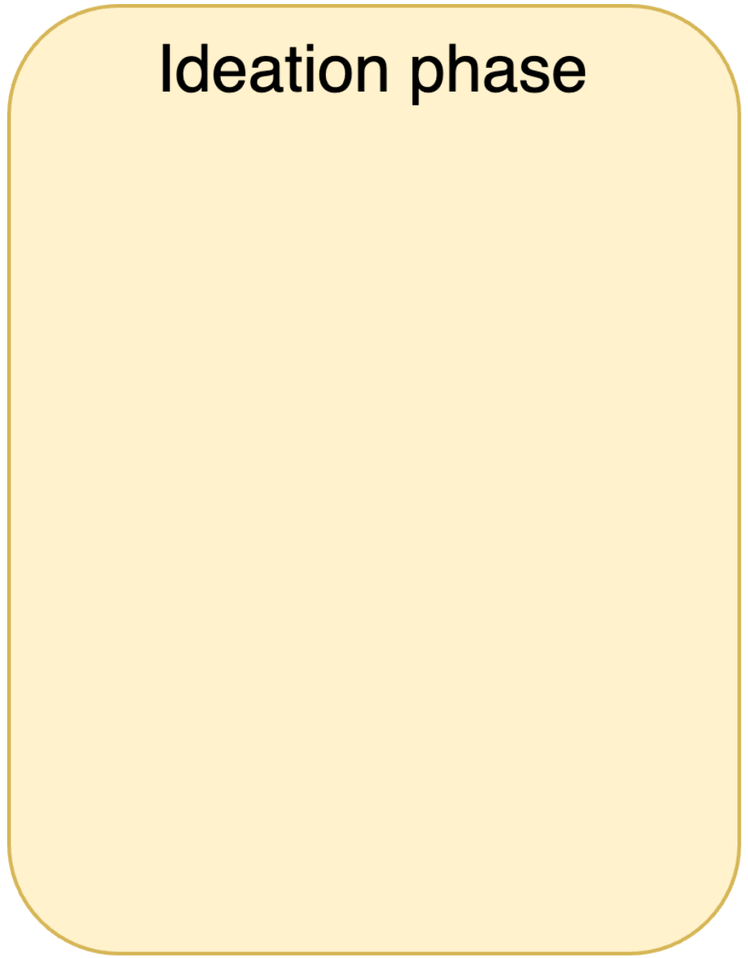
Lifecycle of LLMs

LLMOPS CONCEPTS

Max Knobbout, PhD
Applied Scientist, Uber

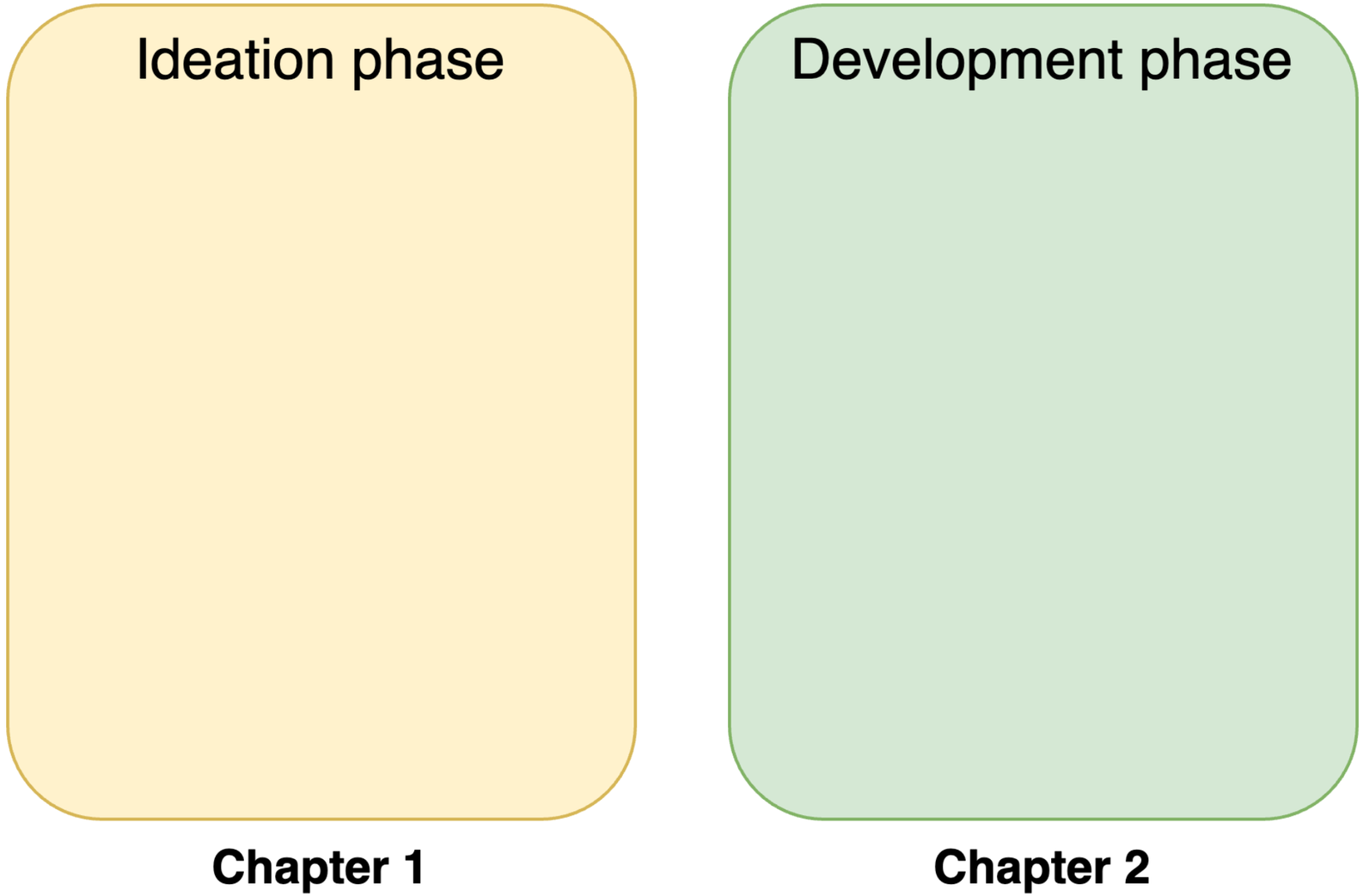
datacamp

The different phases

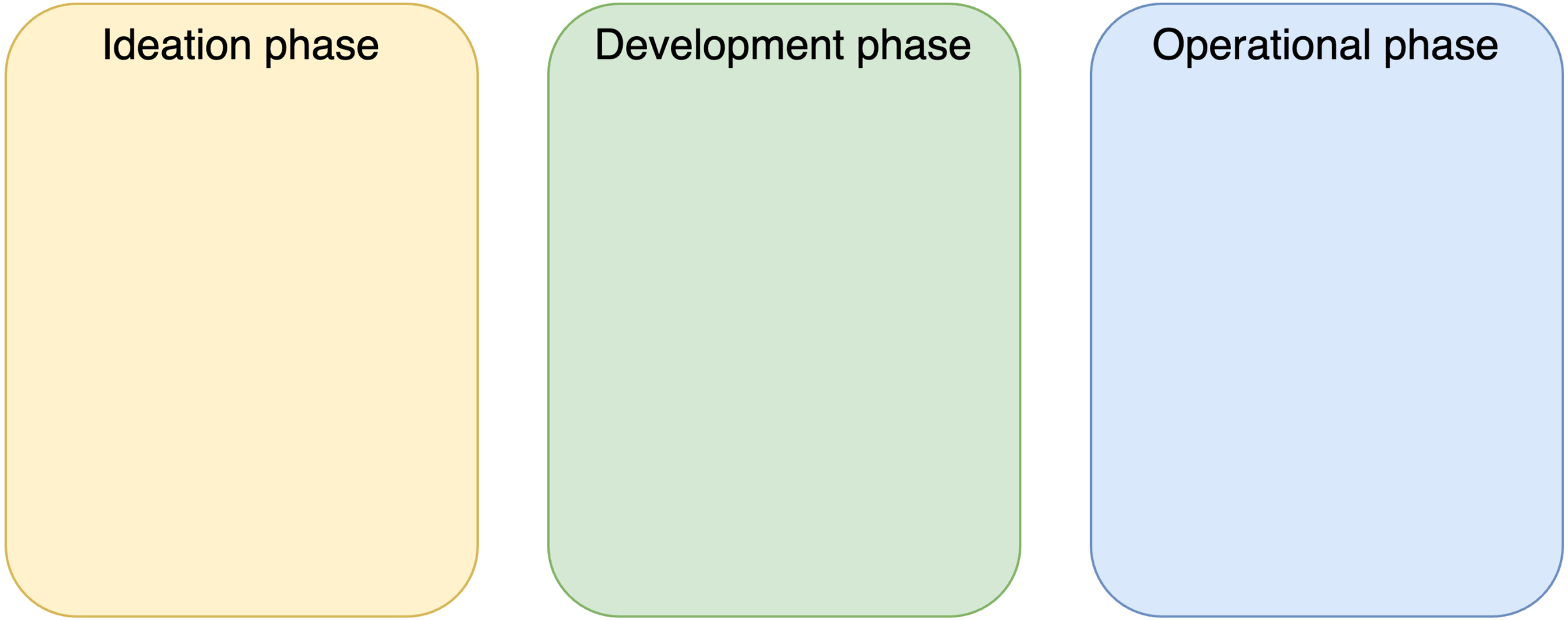


Chapter 1

The different phases

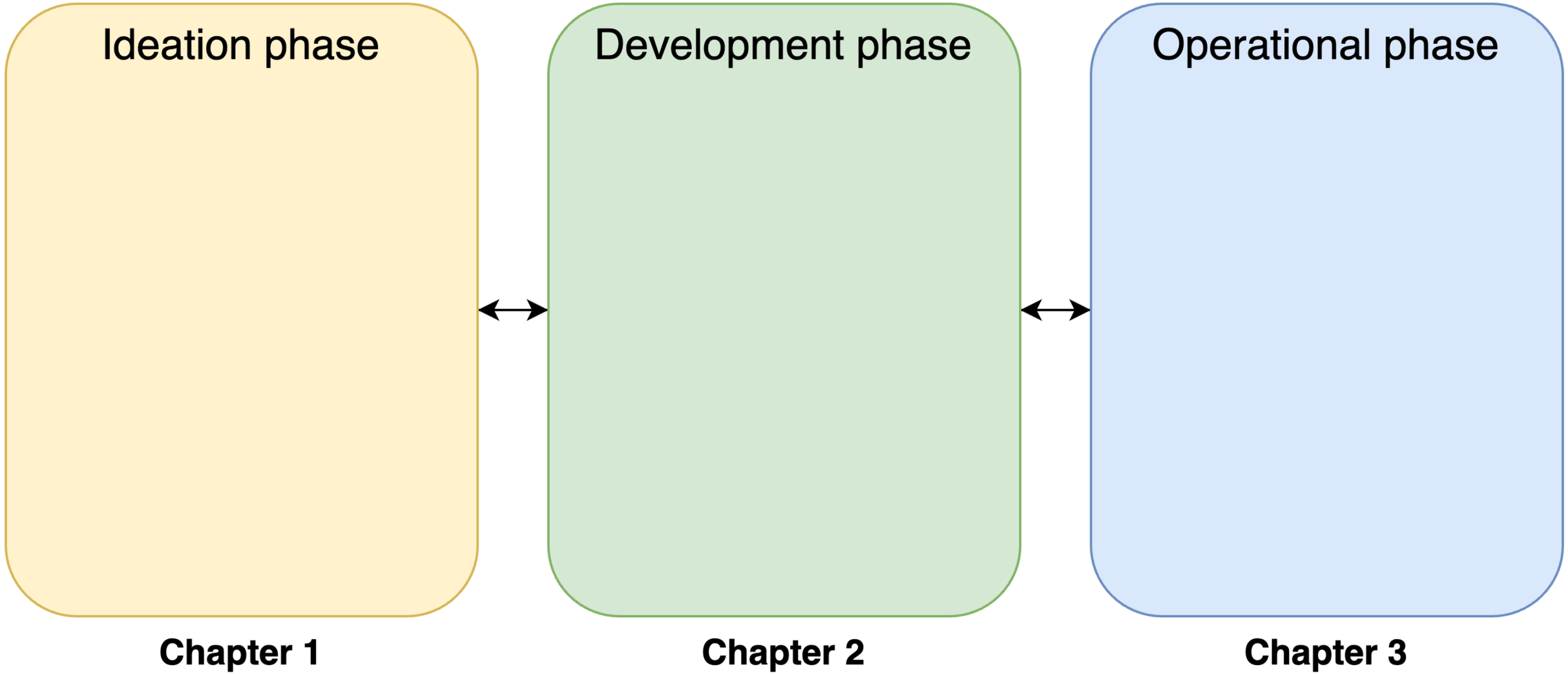


The different phases

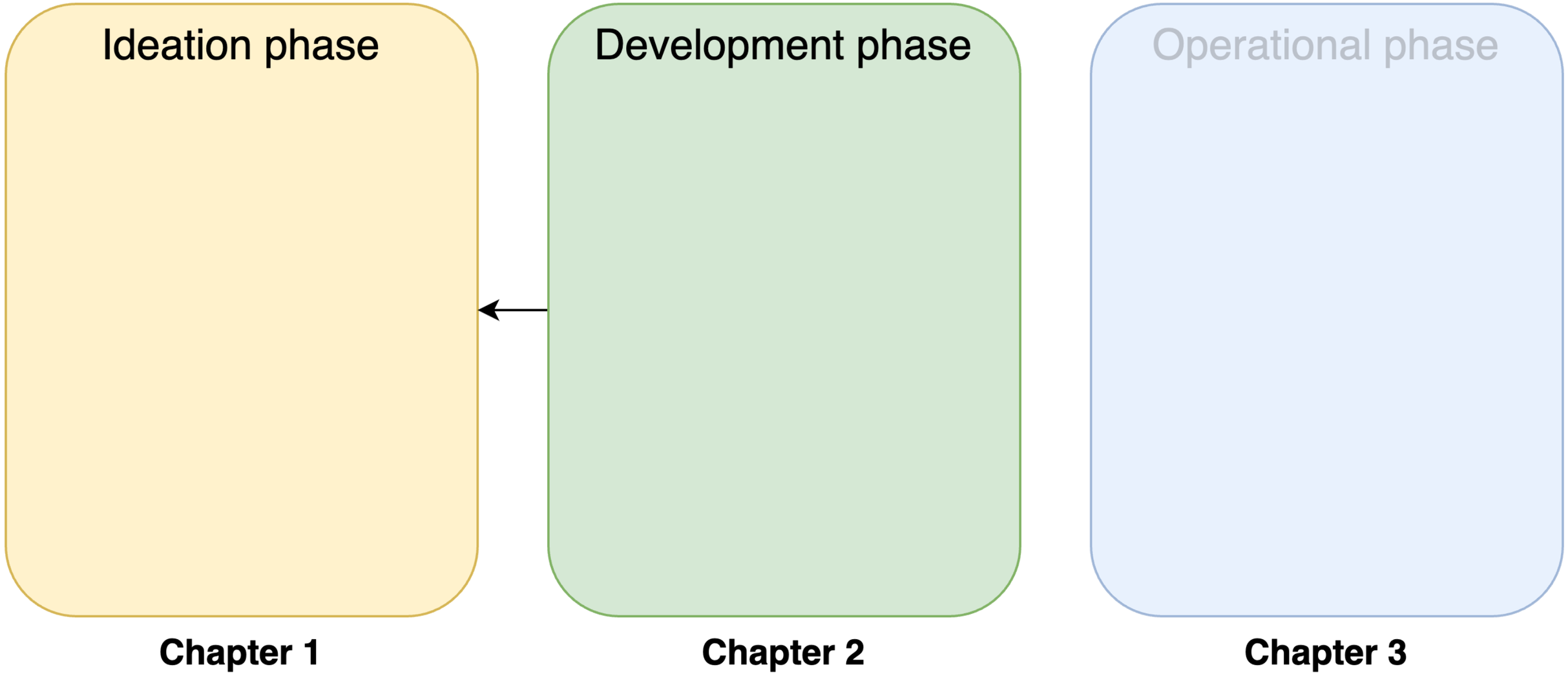


Chapter 1 **Chapter 2** **Chapter 3**

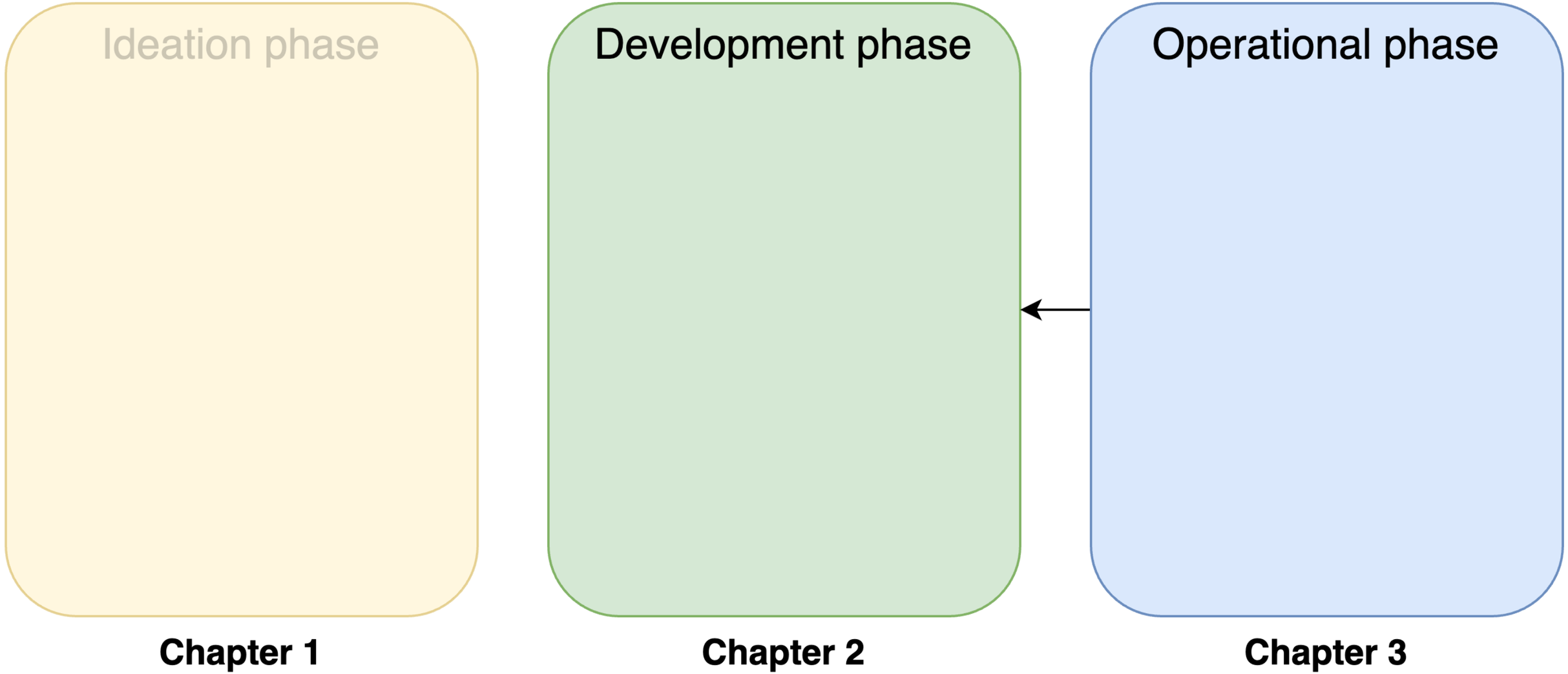
The different phases



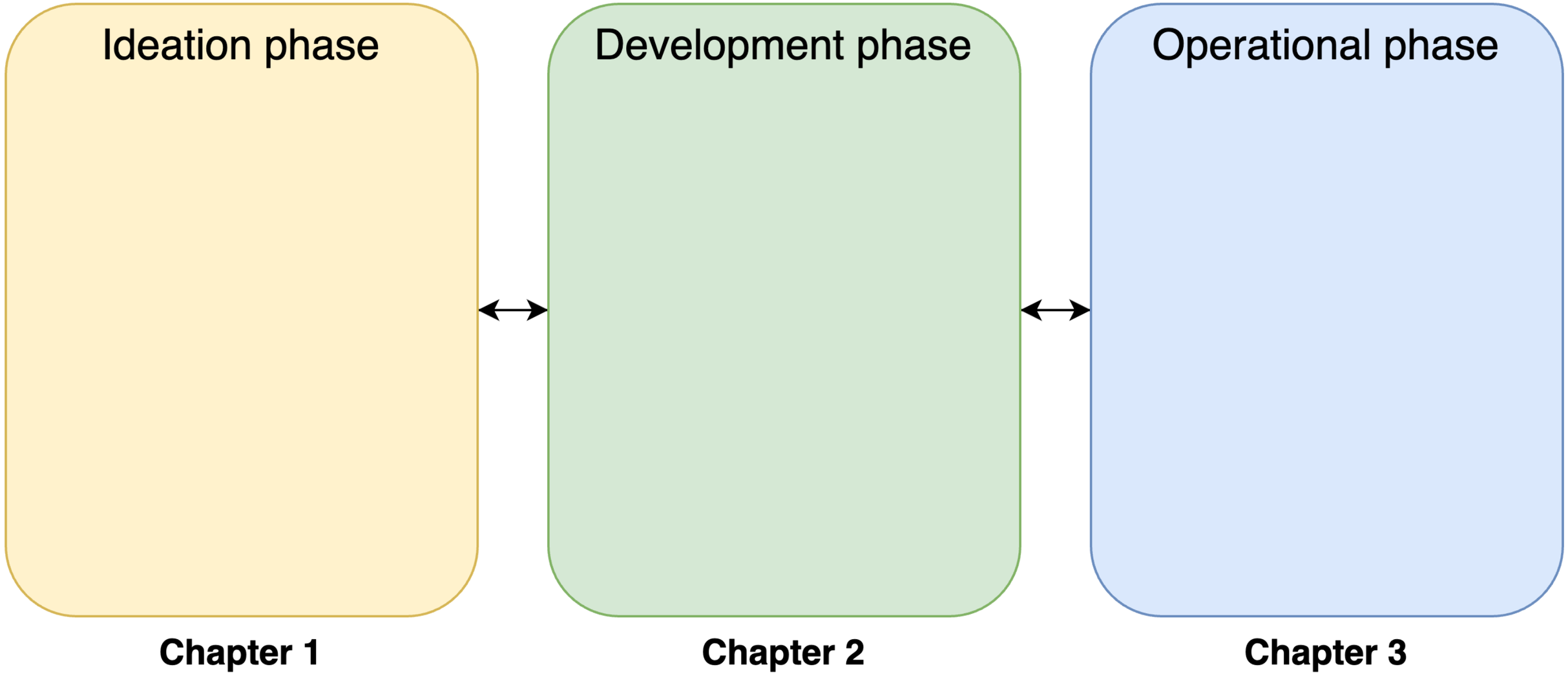
The different phases



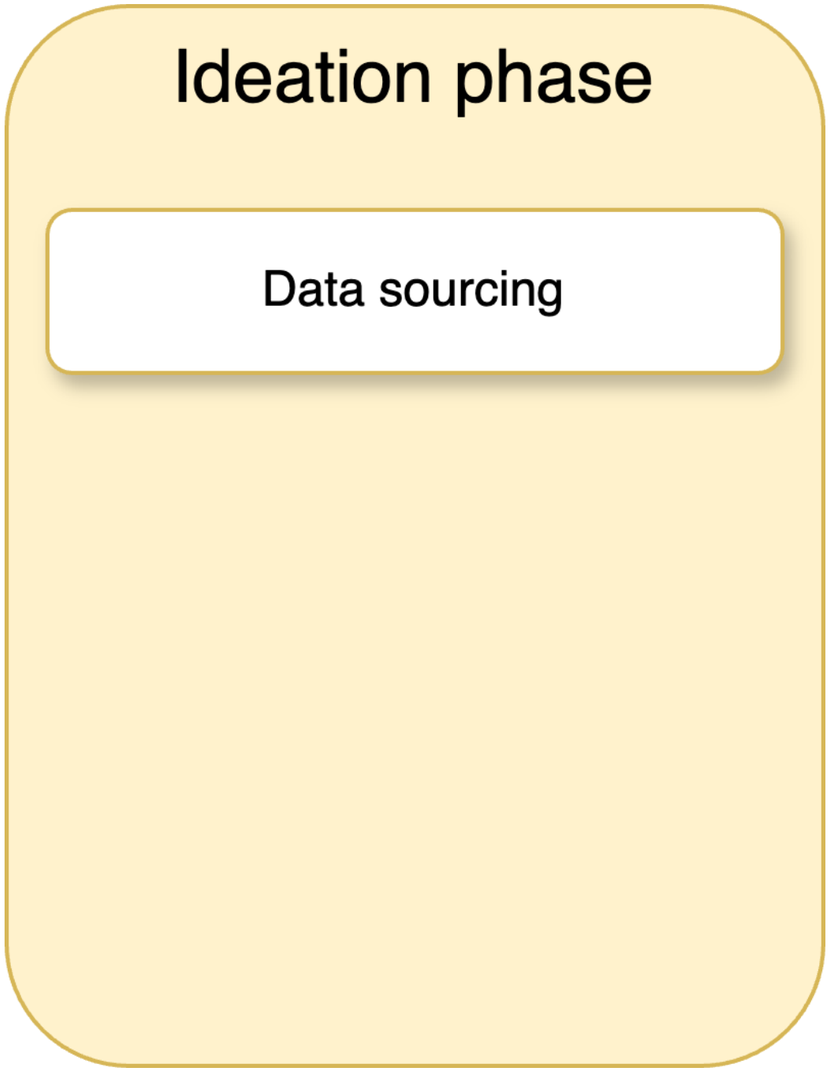
The different phases



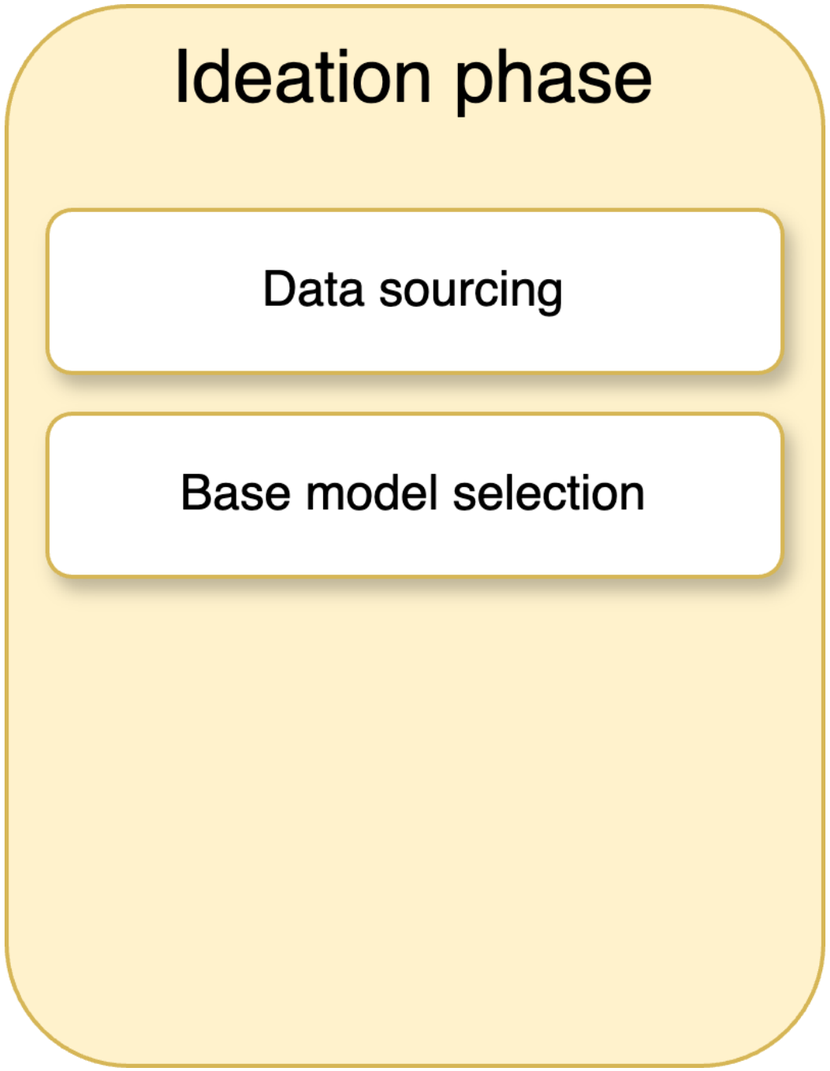
The different phases



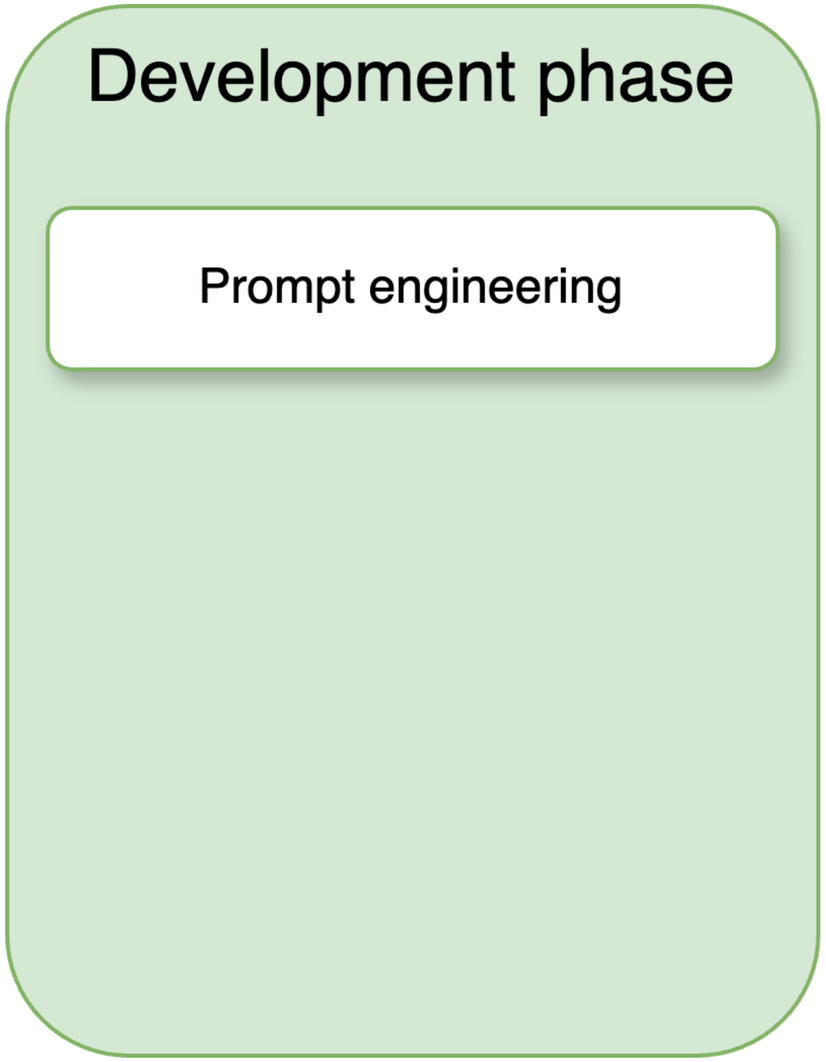
Ideation phase



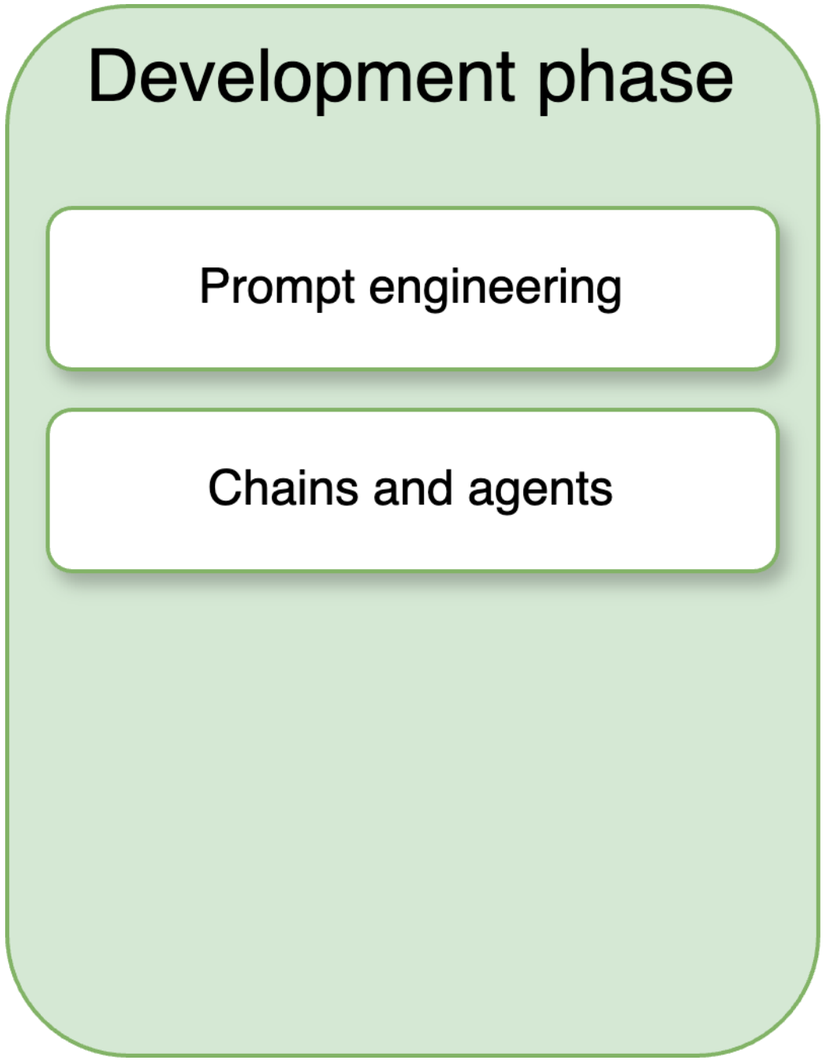
Ideation phase



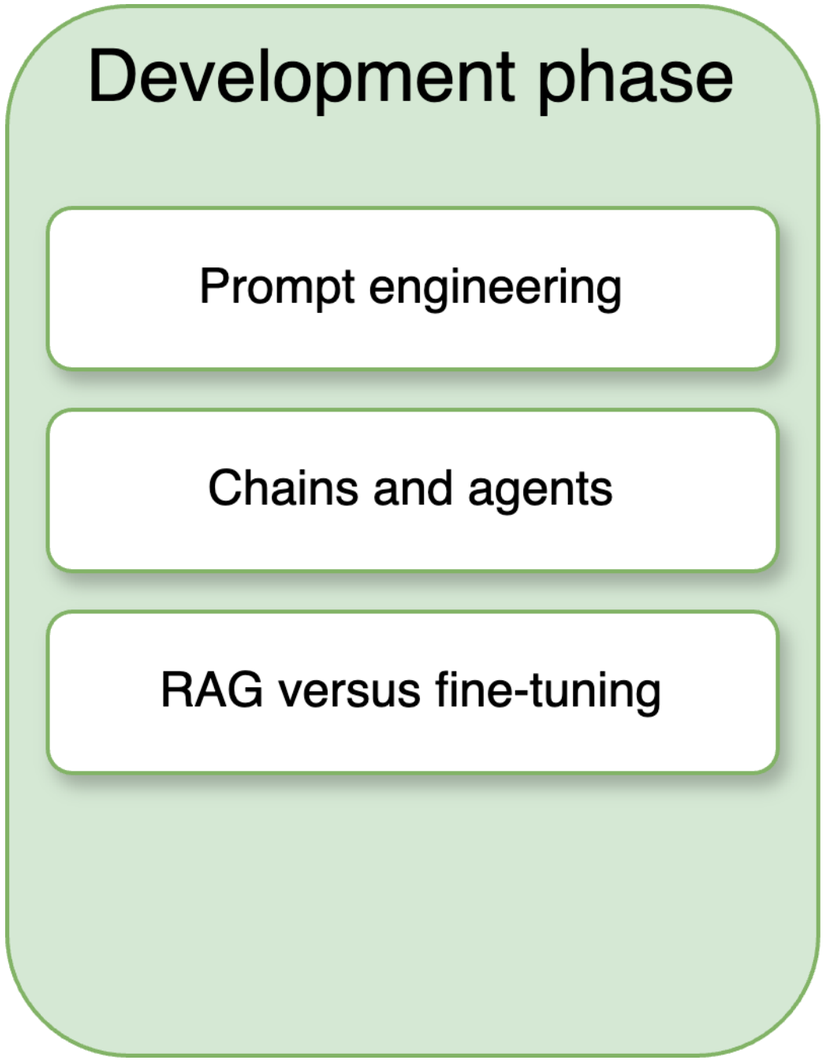
Development phase



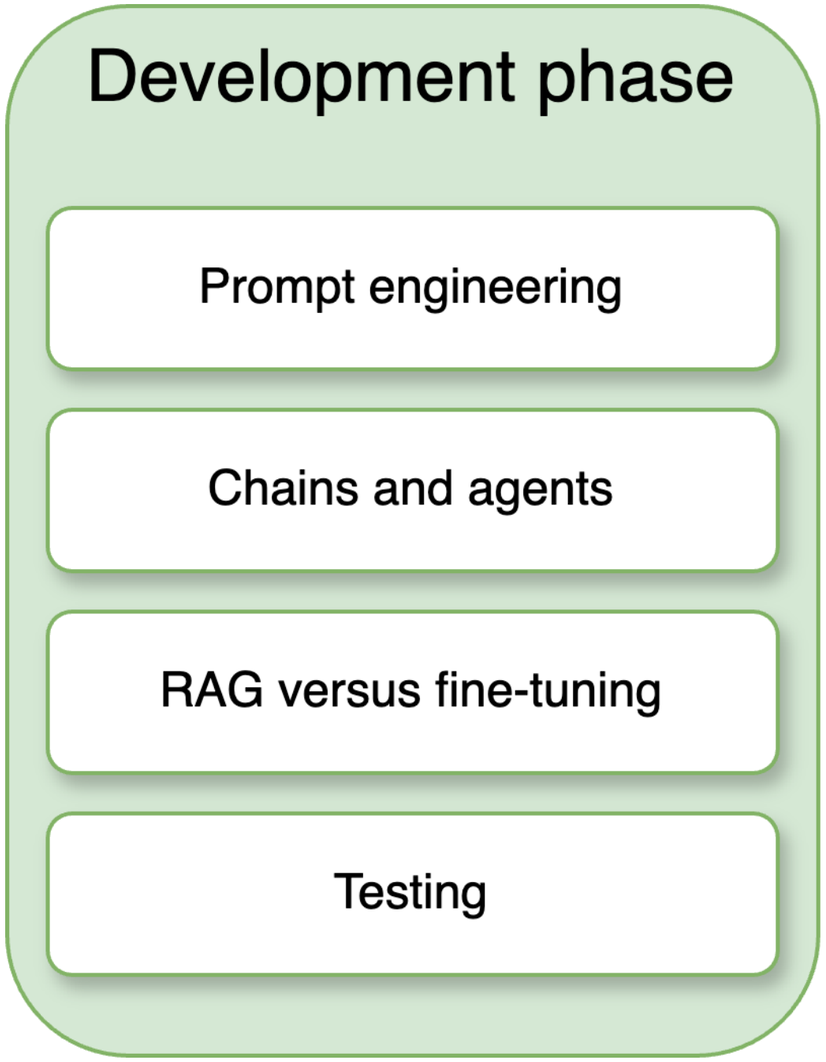
Development phase



Development phase

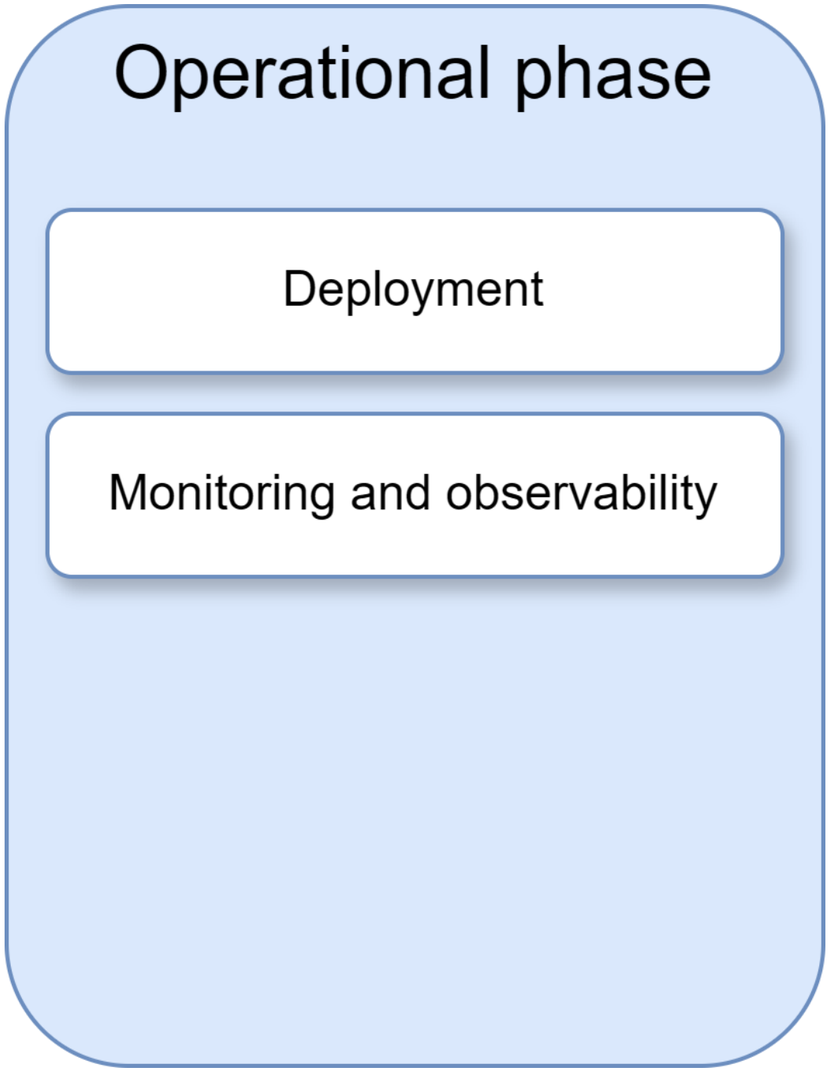


Development phase



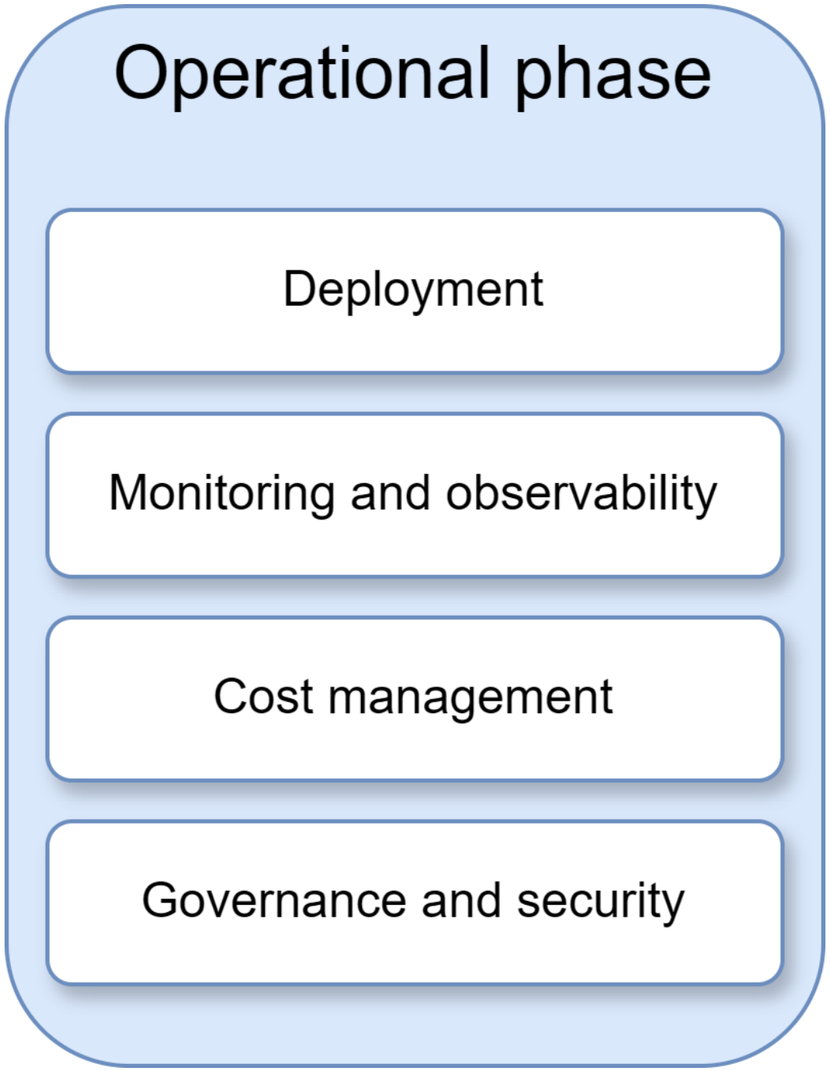
Operational phase

Operational phase

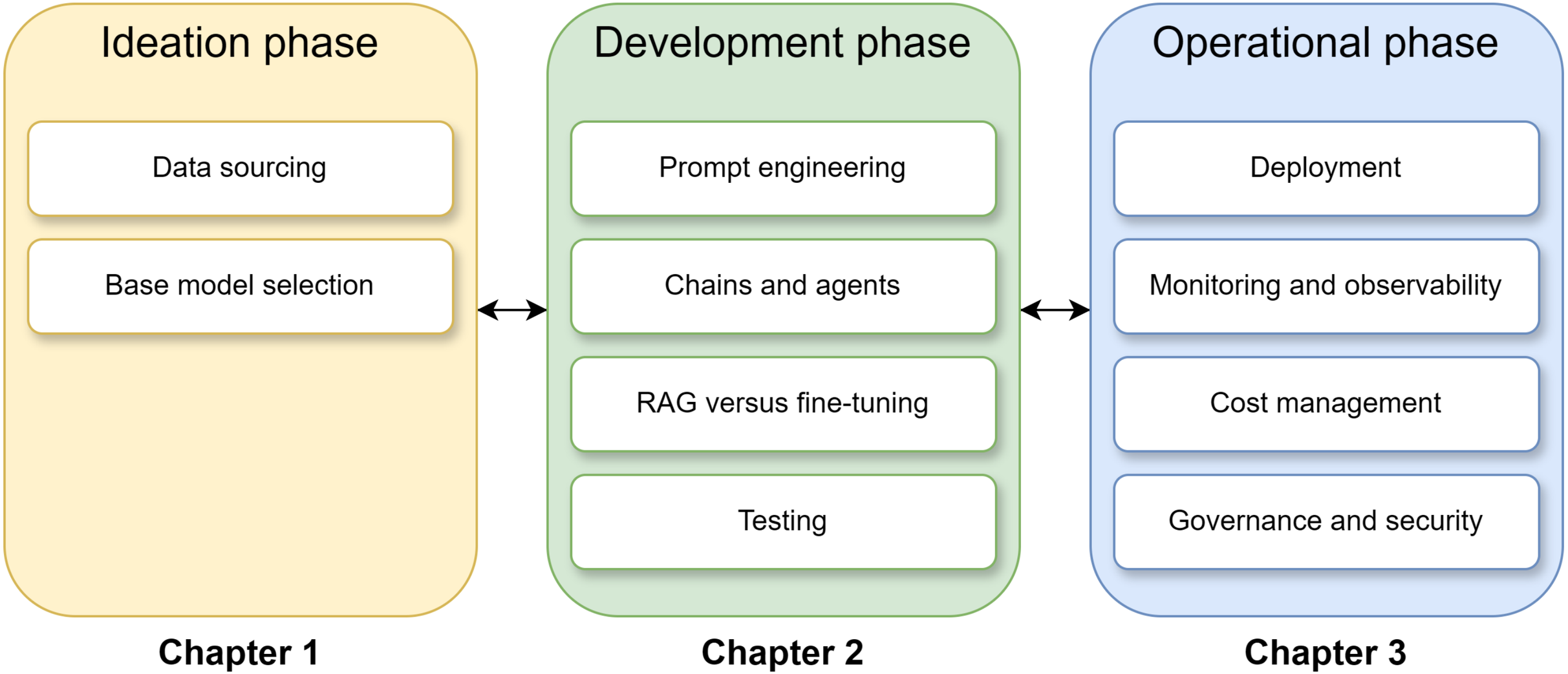


Operational phase

Operational phase



The full picture



Let's practice!

LLMOPS CONCEPTS

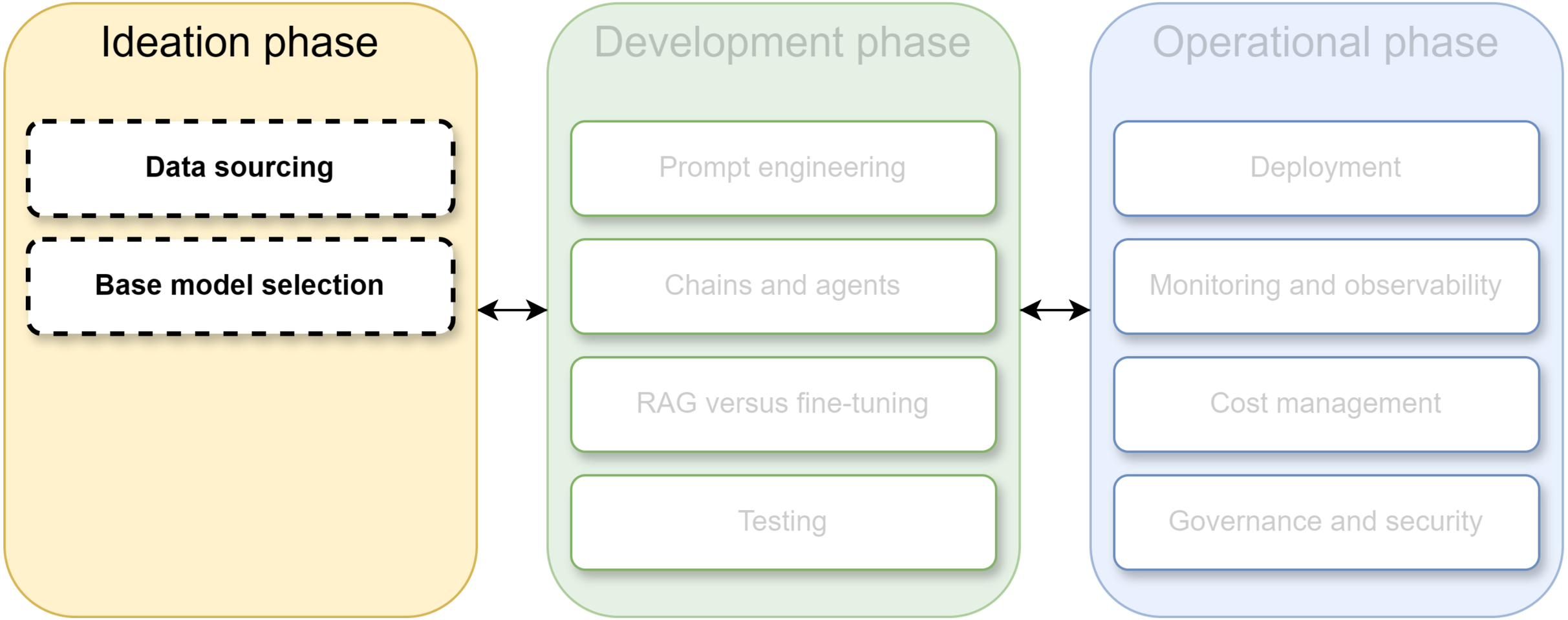
Ideation phase

LLMOPS CONCEPTS

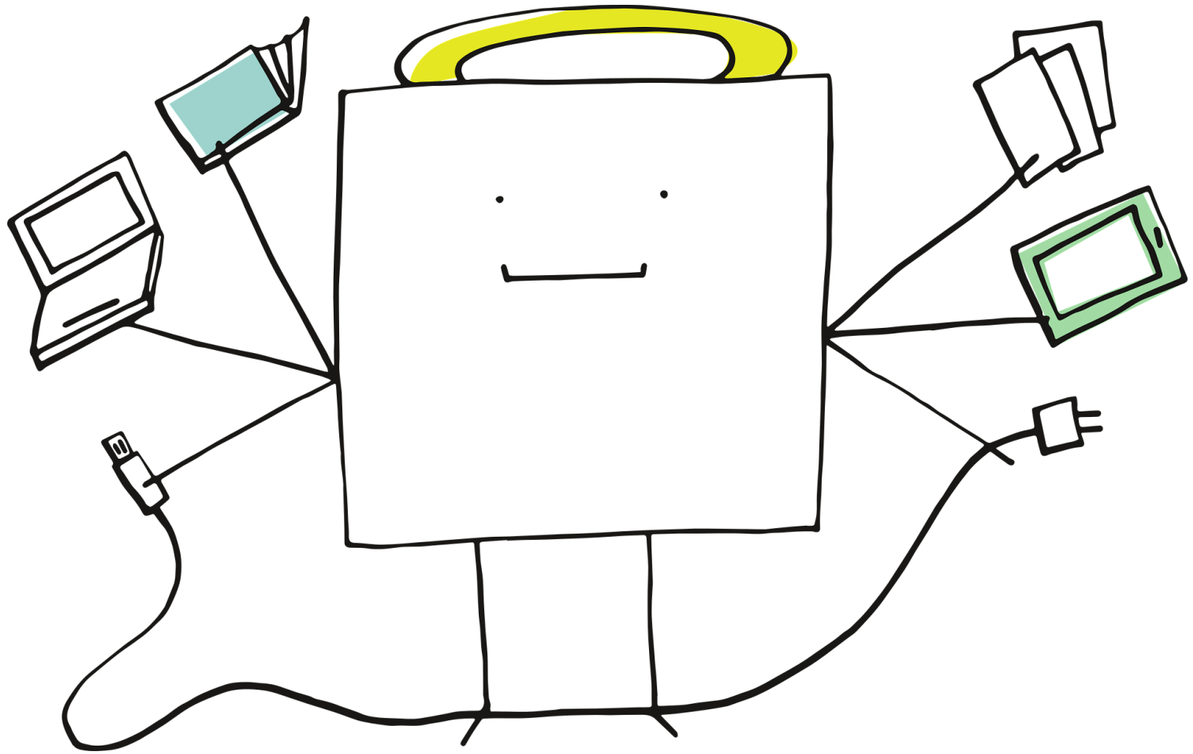
Max Knobbout, PhD
Applied Scientist, Uber

datacamp

LLM lifecyle: Ideation phase

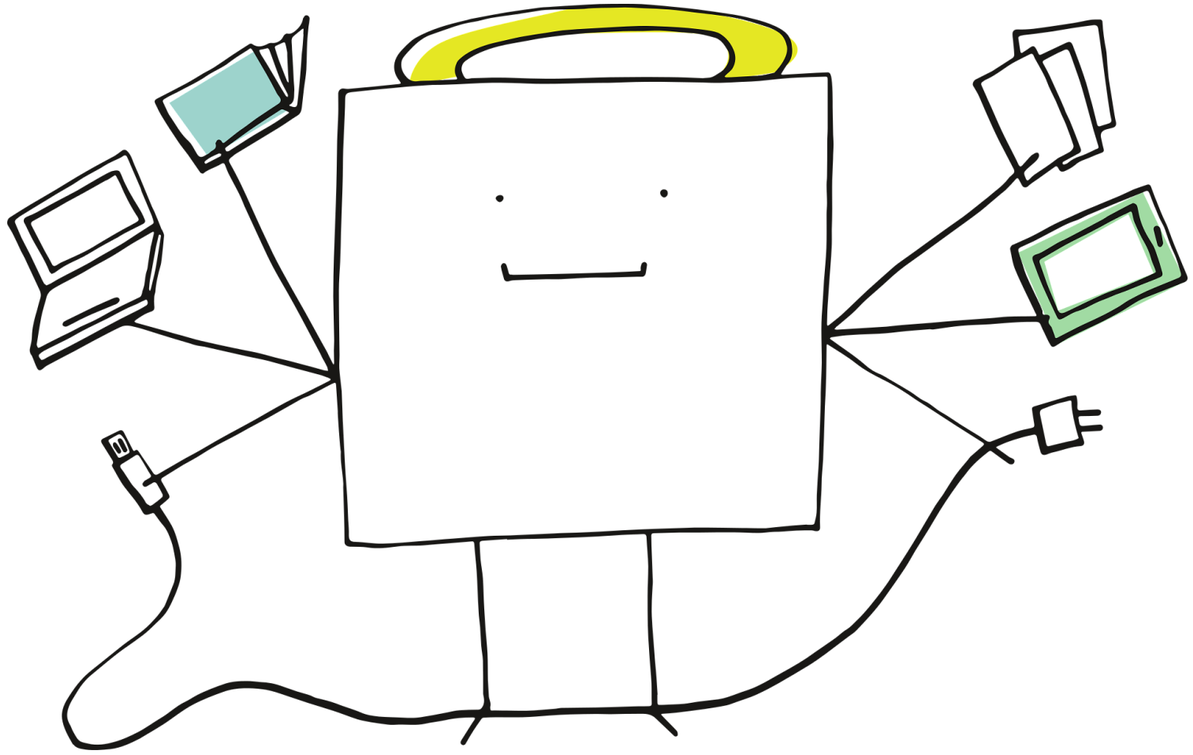


Data sourcing



- Identifying needs
- Finding sources
- Ensuring accessibility

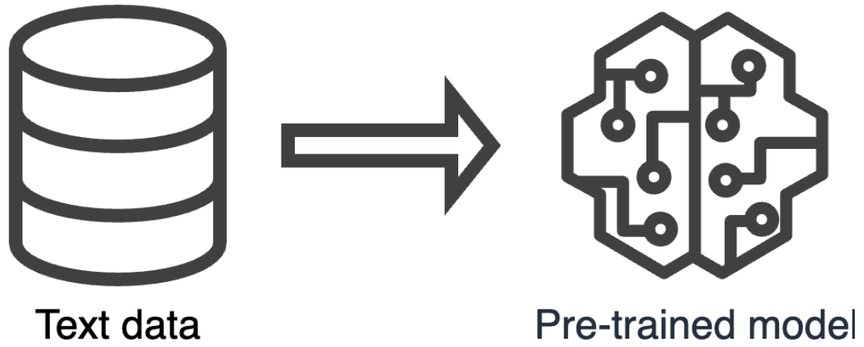
Data sourcing



1. **Is the data relevant?**
2. **Is the data available?**
   - Transform the data
   - Set up additional databases
   - Evaluate costs
   - Consider other access limitations
3. **Does the data meet standards?**
   - Concerns quality and governance

Selecting the base model

- Pre-trained models:



- Proprietary or open-source?

Proprietary models (privately owned)

Advantages:

- Ease of set-up and use
- Quality assurance
- Reliability, speed, and availability

Limitations:

- Requires exposing data
- Customization

Examples:

ChatGPT

Claude 3

PaLM 2

Gemini

Open-source (publicly accessible)

Advantages:

- In-house hosting
- Transparency
- Full customizability

Limitations:

- Support
- Commercial use

Example:

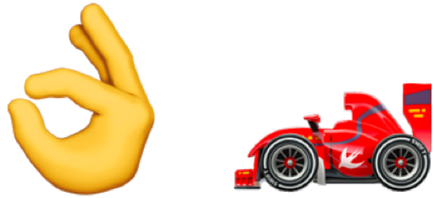
Meta Llama 3

Downloadable from:

Hugging Face

Factors in model selection

1. Performance



- Response quality
- Speed

2. Model Characteristics

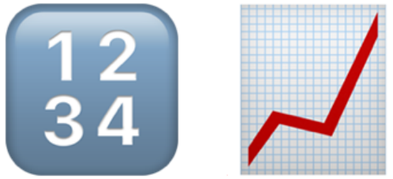
- Data used to train the model
- Context window size
- Fine-tunability

Factors in model selection

3. Practical Considerations

- License
- Cost
- Environmental impact

4. Secondary factors



- Number of parameters
- Popularity

Let's practice!

LLMOPS CONCEPTS