Testing AI Applications

A Whole New Ballgame

Adam Englander

A few questions…

What do you do?

Use end-to-end test automation?

AI application?

Worked with Data Science Teams?

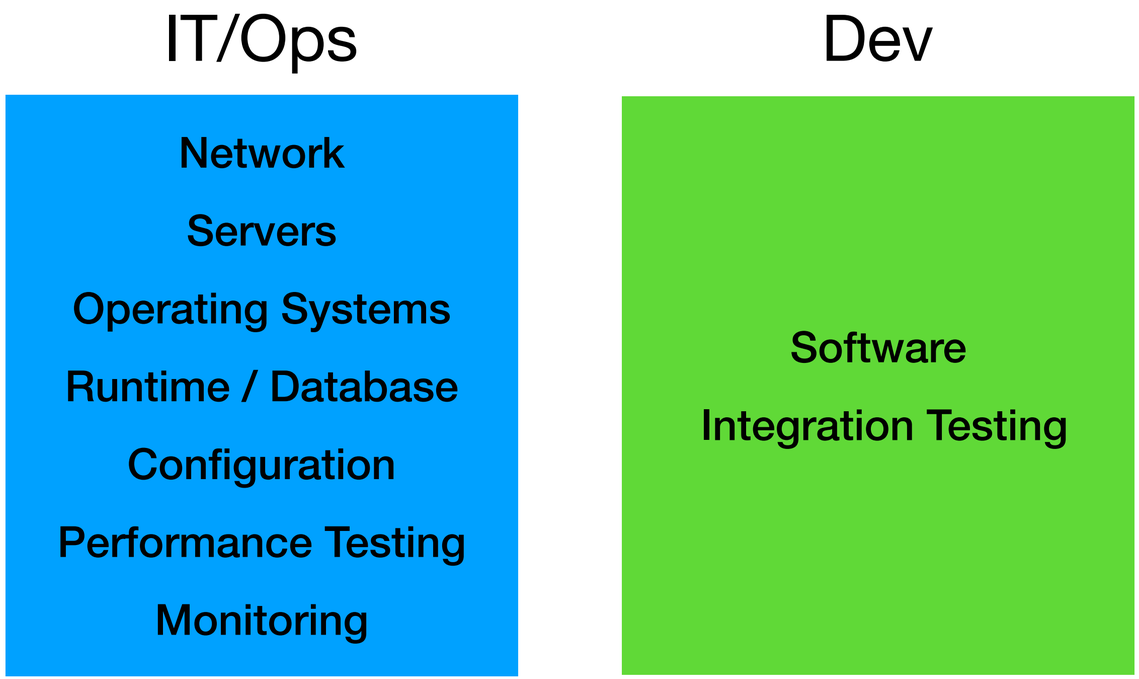
AI is exploding!

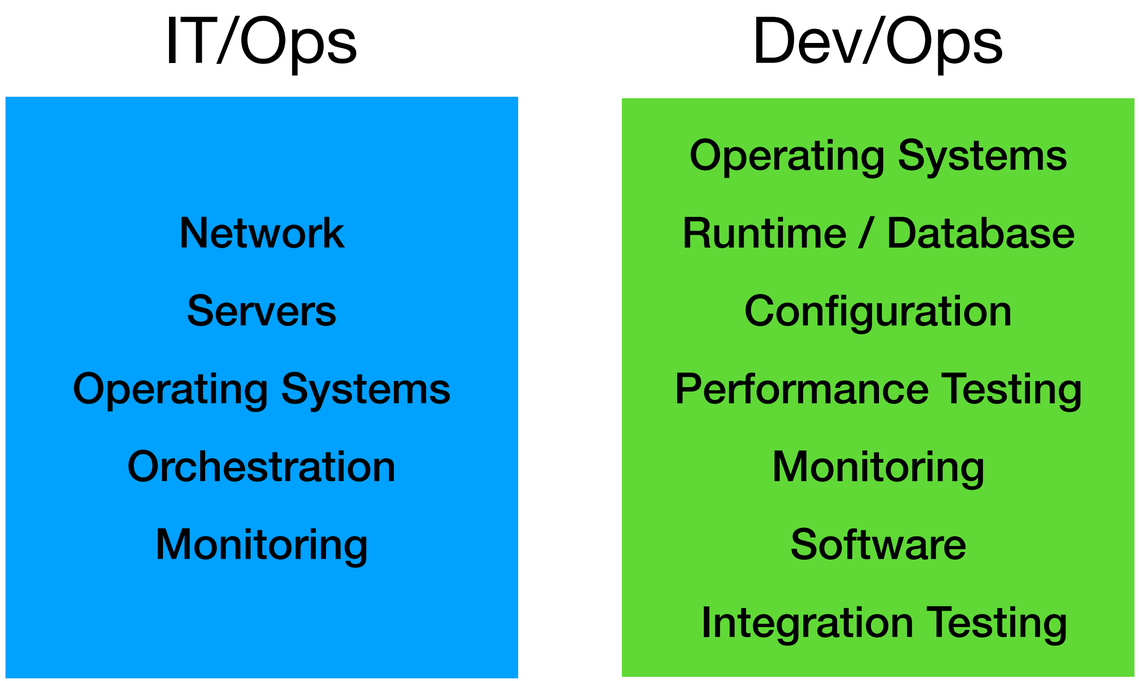
AI is exploding!

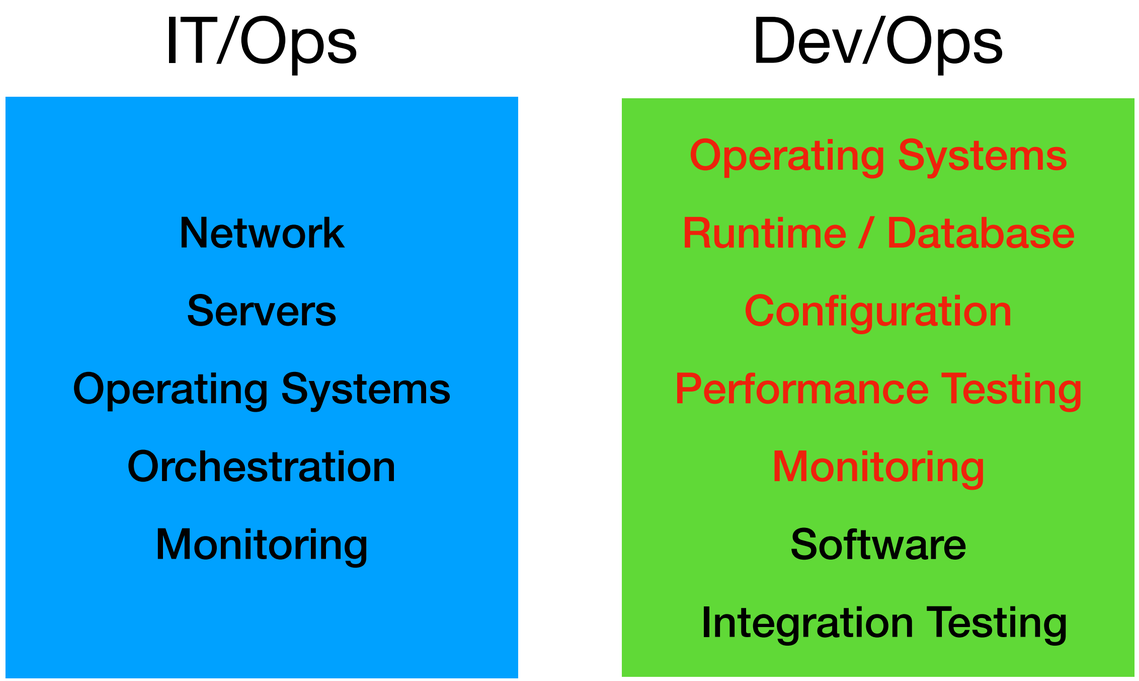
AI is exploding!

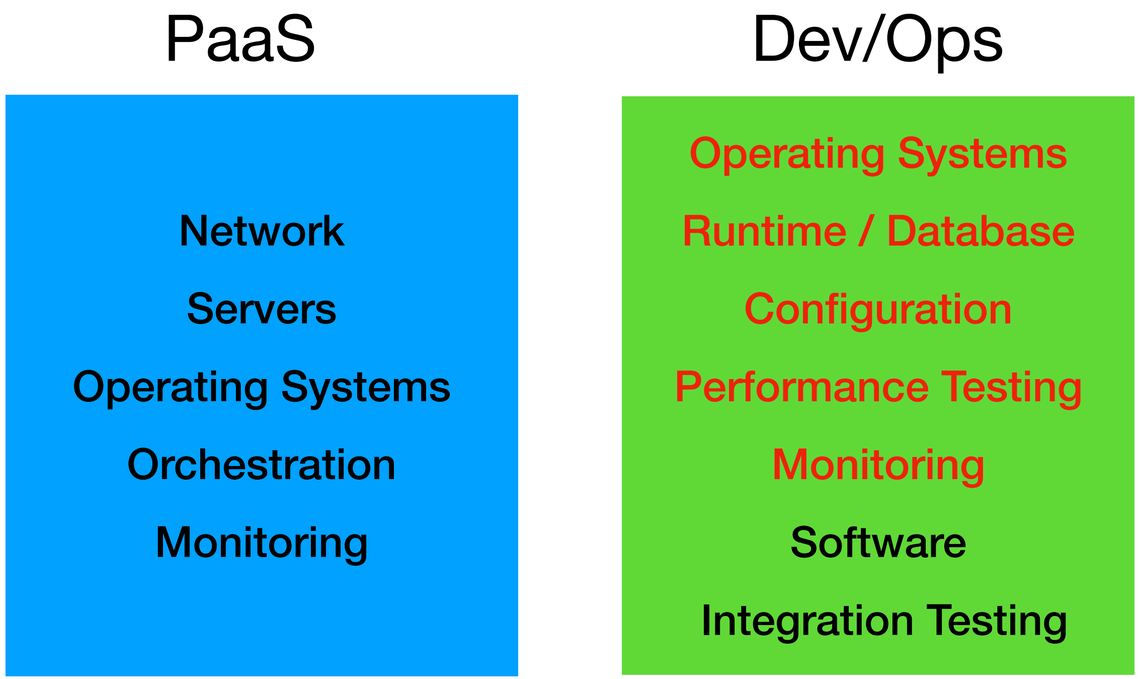
Shifting Responsibility

Shifting Ops Responsibility

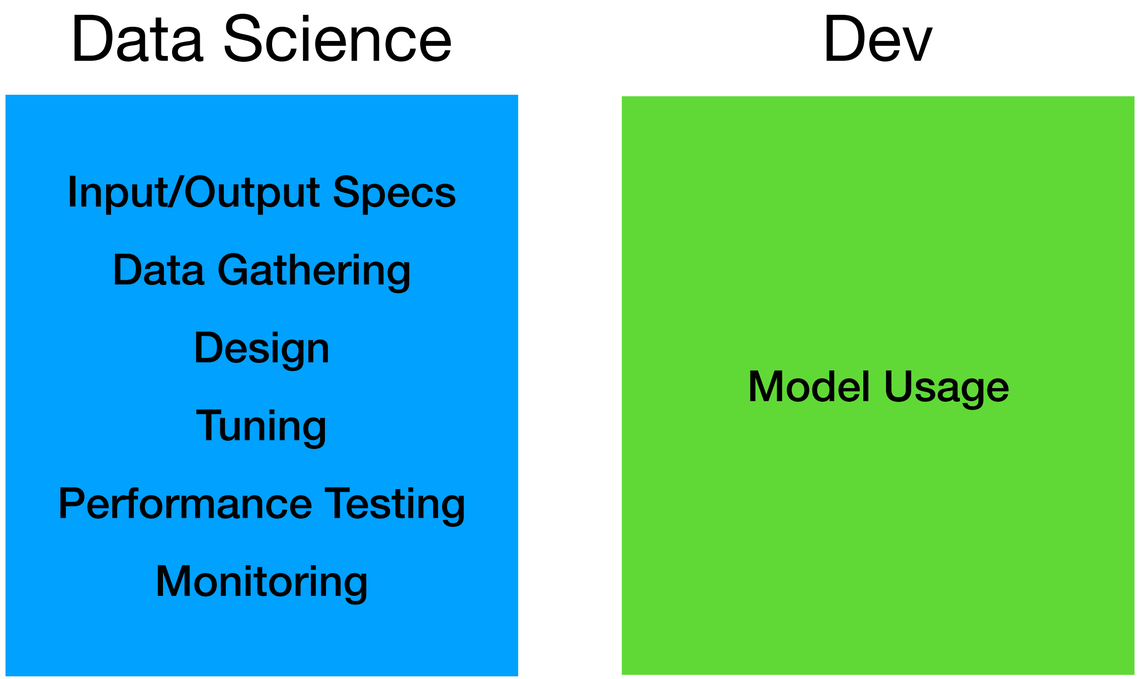

IT/Ops
Network
Servers
Operating Systems
Runtime / Database
Configuration
Performance Testing
Monitoring
Dev
Software
Integration Testing

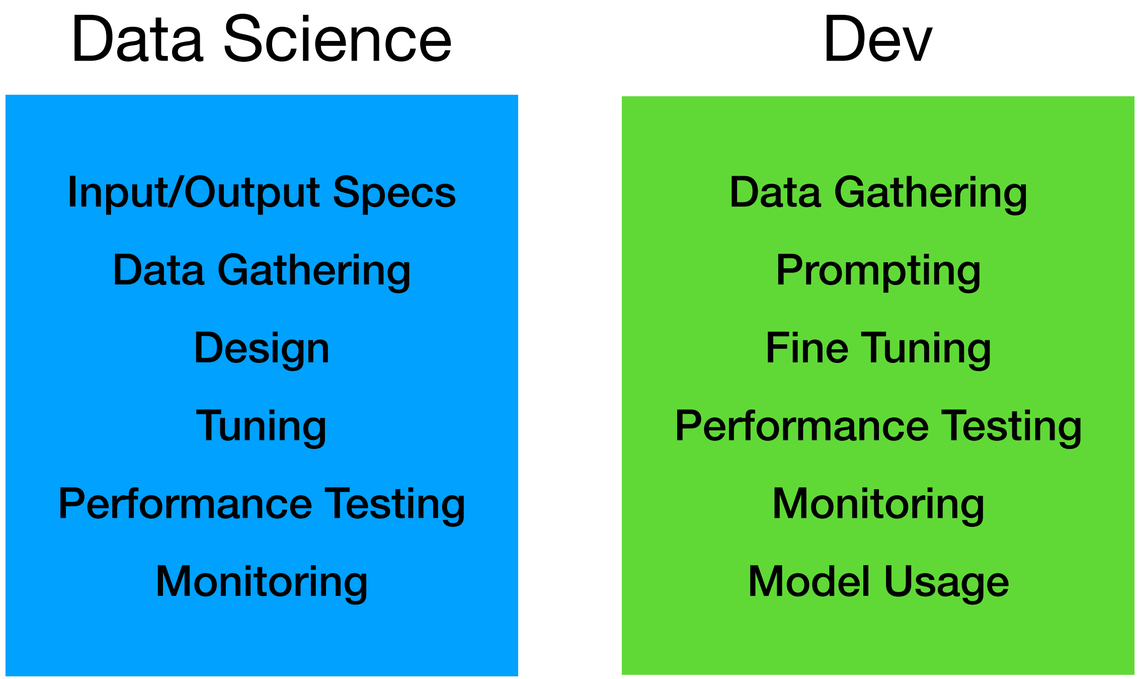

IT/Ops
Network
Servers
Operating Systems
Orchestration
Monitoring
Dev/Ops
Operating Systems
Runtime / Database
Configuration
Performance Testing
Monitoring
Software
Integration Testing


IT/Ops
Network
Servers
Operating Systems
Orchestration
Monitoring
Dev/Ops
Operating Systems
Runtime / Database
Configuration
Performance Testing
Monitoring
Software
Integration Testing


PaaS
Network
Servers
Operating Systems
Orchestration
Monitoring
Dev/Ops
Operating Systems
Runtime / Database
Configuration
Performance Testing
Monitoring
Software
Integration Testing

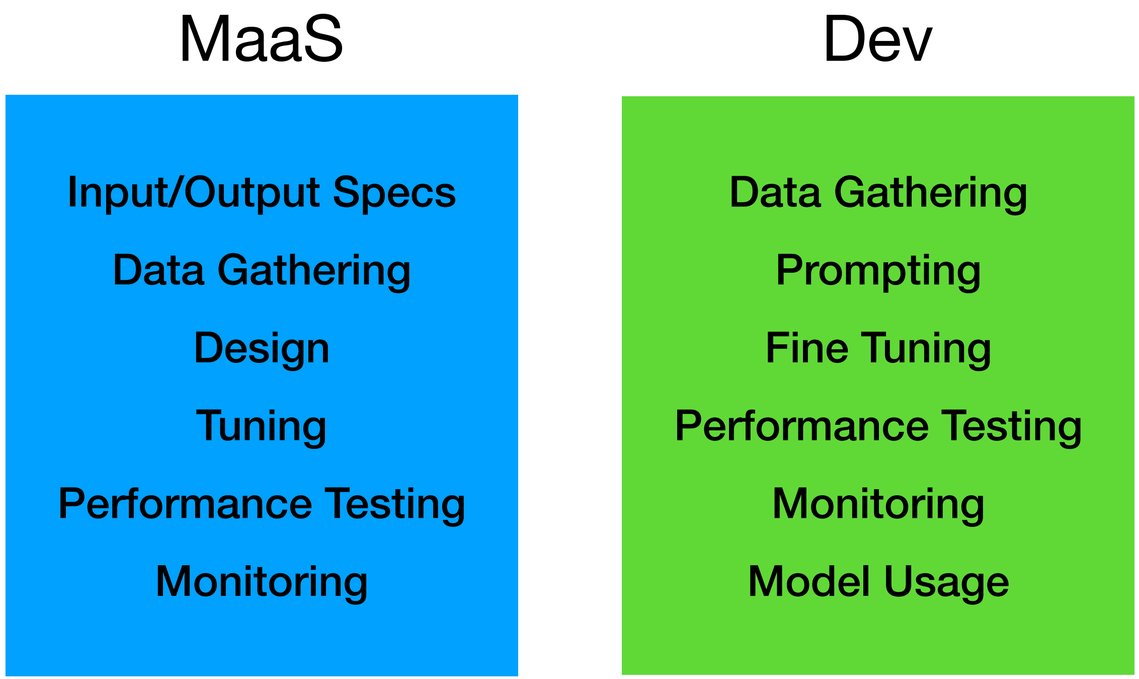
Shifting Data Science Responsibilities


Data Science
Input/Output Specs
Data Gathering
Design
Tuning
Performance Testing
Monitoring
Dev
Model Usage


Data Science
Input/Output Specs
Data Gathering
Design
Tuning
Performance Testing
Monitoring
Dev
Data Gathering
Prompting
Fine Tuning
Performance Testing
Monitoring
Model Usage

MaaS	Dev
Input/Output Specs	Data Gathering
Data Gathering	Prompting
Design	~~Fine Tuning~~
Tuning	Performance Testing
Performance Testing	Monitoring
Monitoring	Model Usage

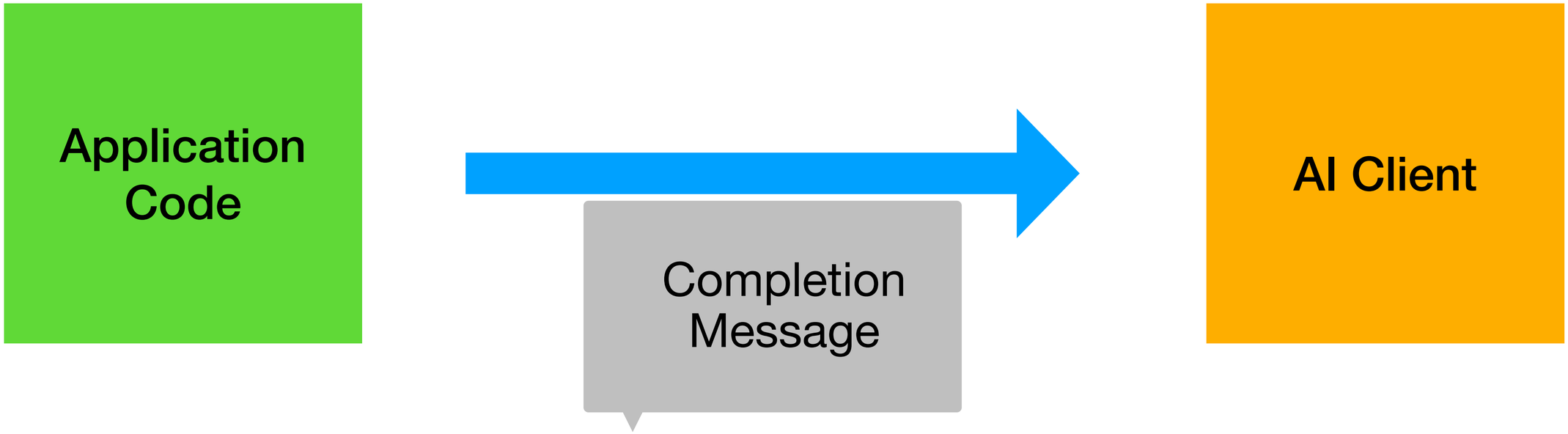
MaaS	Dev
Input/Output Specs	Data Gathering
Data Gathering	Prompting
Design	~~Fine Tuning~~
Tuning	Performance Testing
Performance Testing	Monitoring
Monitoring	Model Usage


MaaS
Input/Output Specs
Data Gathering
Design
Tuning
Performance Testing
Monitoring
Dev
Data Gathering
Prompting
Fine Tuning
Performance Testing
Monitoring
Model Usage

What about vector stores?

How Testing Changes With AI

Unit Testing


Application Code
Completion Message
AI Client


Application Code
Completion Result
AI Client

“What is 1 + 1? Always respond with a valid JSON response.”

- 2
- “2”
- “two”
- {result: 2}
- {“result”: “The result is two.”, “reason”: “The answer was derived by adding the integers one and one together.”
- “A mathematical equation”

Integration and Functional (Feature) Testing

AI Requests = $$$

Fakes/Mocks ≠ $$$

Cheaper Models < $$$

$$A + B \cong C$$

Why is the sky blue?

The sky appears blue because of a phenomenon called Rayleigh scattering.

The sky appears blue because of Rayleigh scattering, which is the scattering of light by gases and particles in Earth's atmosphere.

The sky appears blue because of a phenomenon called Rayleigh scattering.

The sky appears blue because of Rayleigh scattering, which is the scattering of light by gases and particles in Earth's atmosphere.

The sky appears blue because of a phenomenon called Rayleigh scattering.

The sky appears blue because of Rayleigh scattering, which is the scattering of light by gases and particles in Earth's atmosphere.

How do I determine accuracy?

Separate functional from accurate

Separate functional from accurate

Performance Testing

- Request Time
- CPU utilization
- Memory Utilization
- Concurrency
- **Accuracy**

Accuracy is a ratio

The sky appears blue because of a phenomenon called Rayleigh scattering.

The sky appears blue because of Rayleigh scattering, which is the scattering of light by gases and particles in Earth's atmosphere.

The sky appears blue because of a phenomenon called Rayleigh scattering.

The sky appears blue because of Rayleigh scattering, which is the scattering of light by gases and particles in Earth's atmosphere.

The sky appears blue because of a phenomenon called Rayleigh scattering, which the scattering of light by gases derived from Smurf flatulence which is captured in the Earth's atmosphere.

The ancient Greeks believed the sky is blue due its divine nature representing the purity and calmness of the god Uranus.

How wrong are they?

- Rayleigh scattering
- Rayleigh scattering w/definition
- Smurf flatulence
- Ancient Greek explanation

How wrong are they?

- Rayleigh scattering
- Rayleigh scattering w/definition
- Smurf flatulence
- Ancient Greek explanation

How wrong are they?

- Rayleigh scattering
- Rayleigh scattering w/definition
- Smurf flatulence
- ~~Ancient Greek explanation~~

How wrong are they?

- ~~Rayleigh scattering~~
- Rayleigh scattering w/definition
- ~~Smurf flatulence~~
- ~~Ancient Greek explanation~~

How wrong are they?

- Rayleigh scattering
- Rayleigh scattering w/definition
- ~~Smurf flatulence~~
- Ancient Greek explanation

How often are they wrong?

- Rayleigh scattering w/definition
- Rayleigh scattering
- ~~Smurf flatulence~~
- Greek god Uranus
- ~~Smurf flatulence~~
- ~~Smurf flatulence~~
- ~~Smurf flatulence~~

Performance tests allow developers to determine if model/prompt changes improve or degrade application performance.

Performance tests allow developers to determine if model/prompt changes improve or degrade application performance.

Performance tests allow developers to determine if **model/prompt changes** improve or degrade application performance.

Performance tests allow developers to determine if **model/prompt changes** improve or degrade application performance.

Performance tests allow developers to determine if **model/prompt changes** improve or degrade application performance.

How do I test?

How do I test?

Rate of Accuracy rather than Pass/Fail

Rate of Accuracy rather than Pass/Fail

Run a performance test as many times as you can afford.

Run a performance test as many times as you can afford.

Boolean Evaluation

Text Similarity Analysis

Text Similarity Analysis

Model-Model Evaluation

Identifying Success Criteria

Item	Weight (1-10)	Scoring Instructions
Grammar	3	Does the response read and flow well?
Accuracy	10	Is the response an accurate representation of a true fact?
Understandable	6	Does the response explain any technical terms it uses?
Relevant	5	Is the response relevant to the question asked?

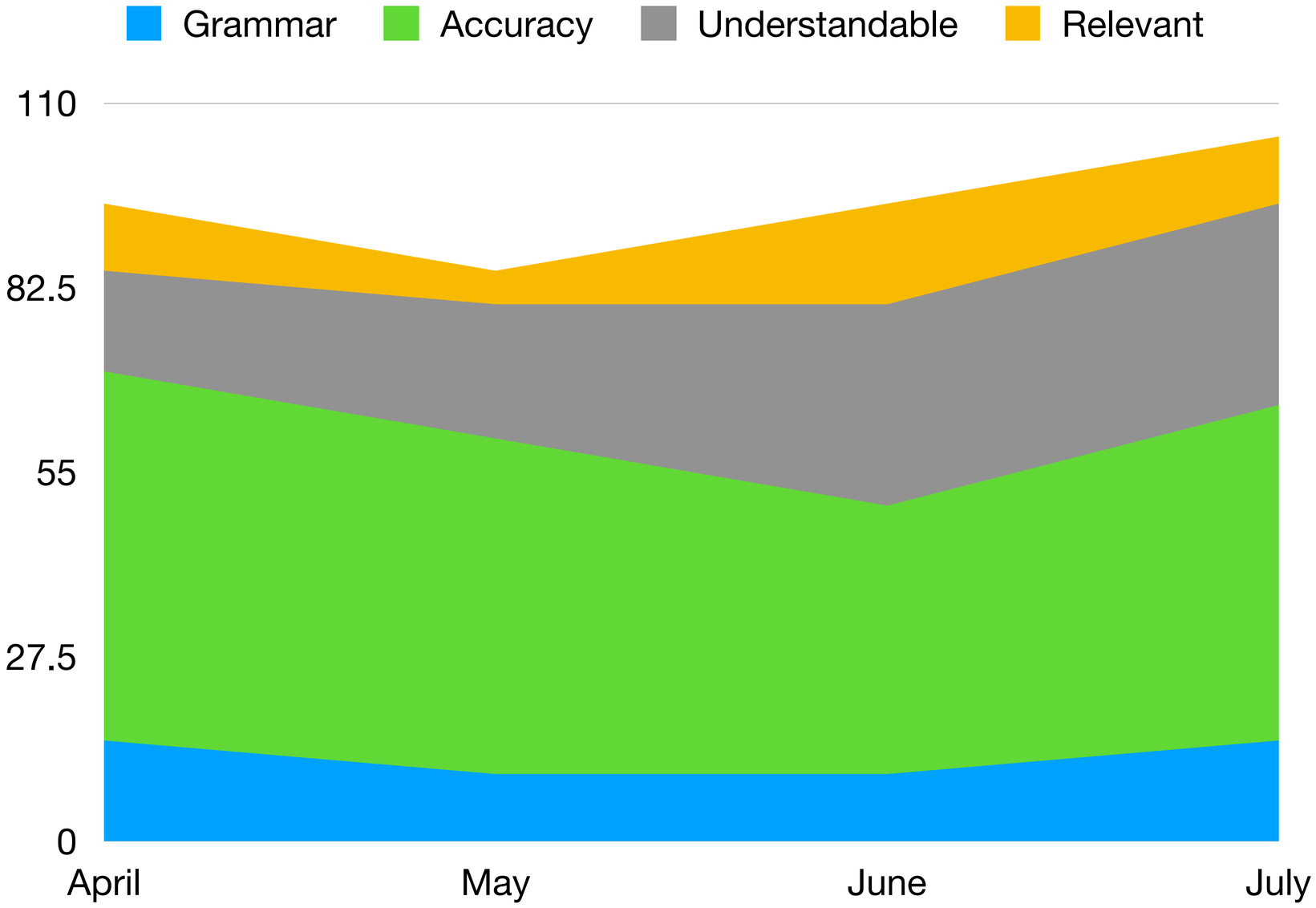
Item	Weight (1-10)	Scoring Instructions
Grammar	3	Does the response read and flow well?
Accuracy	10	Is the response an accurate representation of a true fact?
Understandable	6	Does the response explain any technical terms it uses?
Relevant	5	Is the response relevant to the question asked?

Item	Weight (1-10)	Scoring Instructions
Grammar	**3**	Does the response read and flow well?
Accuracy	**10**	Is the response an accurate representation of a true fact?
Understandable	**6**	Does the response explain any technical terms it uses?
Relevant	**5**	Is the response relevant to the question asked?

Item	Weight (1-10)	Scoring Instructions
Grammar	3	**Does the response read and flow well?**
Accuracy	10	**Is the response an accurate representation of a true fact?**
Understandable	6	**Does the response explain any technical terms it uses?**
Relevant	5	**Is the response relevant to the question asked?**

Accuracy over Time

The accuracy of the results from your model can change without any change to the model or prompts


Grammar
Accuracy
Understandable
Relevant
110
82.5
55
27.5
0
April
May
June
July

User input can change

User input can change

User input can change

Model input can change over time

You will change model versions!

You will change model versions!

In Summary

Verify you’re prepared for unexpected results in unit testing

Reduce development costs by using fakes/mocks where possible and cheaper models where not

Define accuracy and value in your application and create a testing scheme to evaluate both

Validate the accuracy and value of every release

Validate the accuracy and value over time

Feedback

https://confoo.ca/en/2025/feedback/9063F653A5BE18833F6AAC706CEAFF75