DeepLearning.AI

Agentic AI

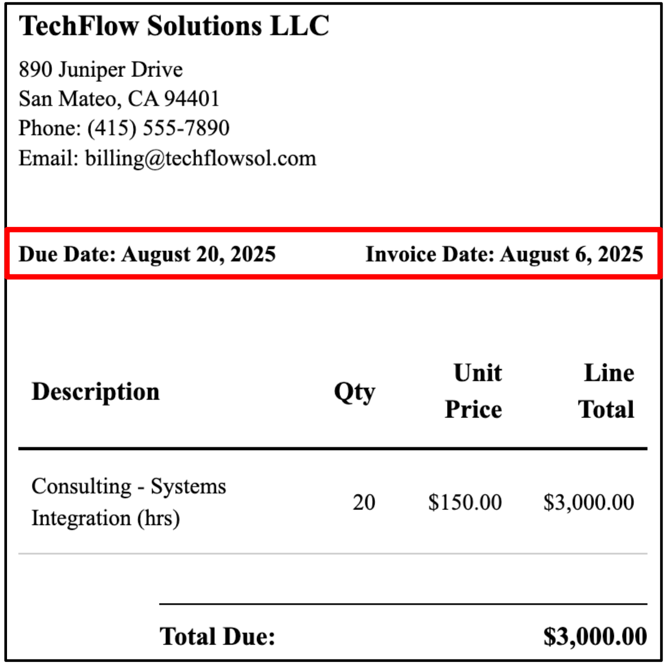
M4: Practical tips for building agentic AI

Practical Tips for Building Agentic AI

Evaluations (evals)

DeepLearning.AI

Example: invoice processing workflow



TechFlow Solutions LLC

890 Juniper Drive
San Mateo, CA 94401
Phone: (415) 555-7890
Email: billing@techflowsol.com

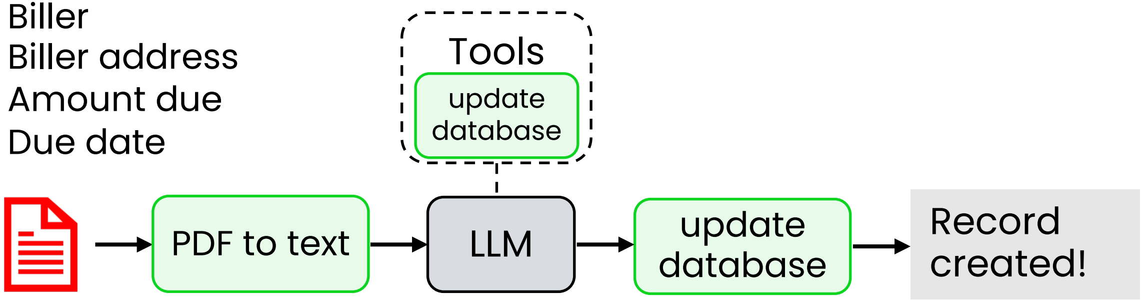
Due Date: August 20, 2025 **Invoice Date: August 6, 2025**

Description	Qty	Unit Price	Line Total
Consulting - Systems Integration (hrs)	20	$150.00	$3,000.00

Total Due: **$3,000.00**

4 required fields:

Biller
Biller address
Amount due
Due date



10–20 invoices

Create an eval to measure date extraction

1. Manually extract due dates from 10-20 invoices

test invoice 1 per example ground truth

"August 20, 2025" → “2025/08/20”

2. Specify output format of data in prompt

Format the due date as YYYY/MM/DD

3. Extract date from the LLM response using code

```
date_pattern = r'\d{4}/\d{2}/\d{2}’
extracted_date = re.findall(date_pattern, llm_response)
```

4. Compare LLM result to ground truth

```
if (extracted_date == actual_date):
    num_correct +=1
```

Driving your development process with evals

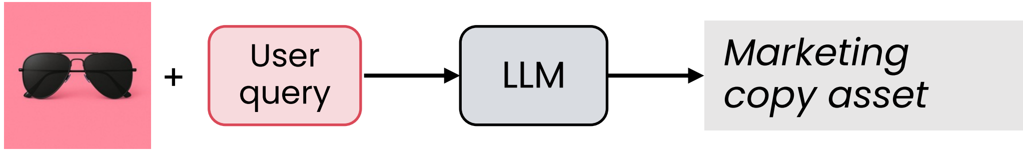
- Build a system and look at outputs to discover where it is behaving in an unsatisfactory way
  - E.g. incorrect due dates in invoice data extract

- Drive improvement by putting in place a small eval with ~20 examples to help you track progress

- Monitor as you make changes to workflow (e.g. new prompts, new algorithms) and see if the metric improves

Example: marketing copy assistant

Length guidelines:
Instagram caption: 10 words max



	17 words
	Ok
	Ok
	14 words
	11 words

Create an eval to measure text length

1. Create a set of 10-20 test tasks

Image	Example prompt
	Create an Instagram post
	Create an Instagram post

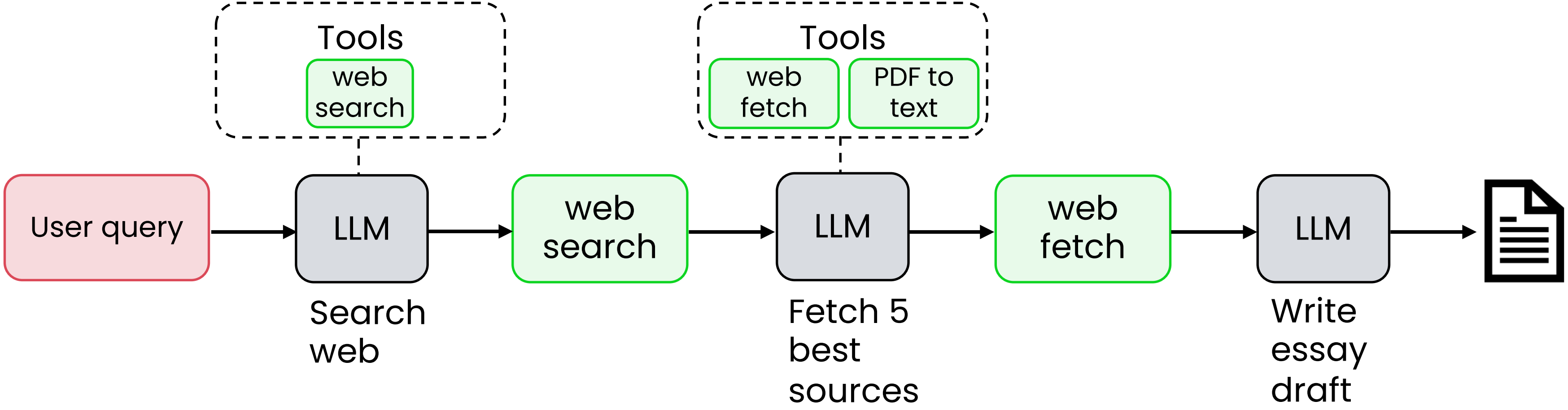
2. Add code to measure word count of the output

```
word_count = len(text.split())
```

3. Compare length of generated text to limit

```
if (word_count <= 10):
     num_correct +=1
```

Example: research agent



Prompt	Issues
Recent black hole science	Missed high-profile result that had lots of news coverage
Renting vs buying a home in Seattle?	Seems to do a good job
Robotics for harvesting fruit	Didn’t mention leading equipment company

Sometimes misses points a human would have made

Create an eval to measure performance

1. Choose 3-5 gold standard discussion points for each topic
2. Use LLM-as-a-judge to count how many topics were mentioned
3. Get score for each prompt in eval set

Example prompt	Gold-standard talking points
Black holes	Event horizon, radio telescope
Robotic harvesting	RoboPick, pinchers

ground truth annotations

Determine how many of the 5 gold-standard talking points are present in the provided essay.

Original Prompt

{original_prompt}

Essay to Evaluate

{essay_text}

Gold Standard Talking Points

{gold_standard_points}

Output Format

Return a json object with two keys: score (a single number between 0 and 5), and explanation (a string that lists the talking points present)

Two “axes” of evaluation

<table>
<tr><th></th><th>Evaluate with code (objective)</th><th>LLM-as-judge (subjective)</th></tr>
<tr><td>Per example ground truth</td><td>Checking invoice date extraction
<pre>
if (extracted_date == actual_date):
    num_correct +=1
</pre></td><td>Counting gold-standard talking points
Count the number of gold standard points in the following text…</td></tr>
<tr><td>No per example ground truth</td><td>Checking marketing copy length
<pre>
if len(text) <= 10:
    num_correct += 1
</pre></td><td>Grading charts with a rubric
Grade this chart according to (i) whether it has clear axes labels, (ii) ….</td></tr>
</table>

Tips for designing end-to-end evals

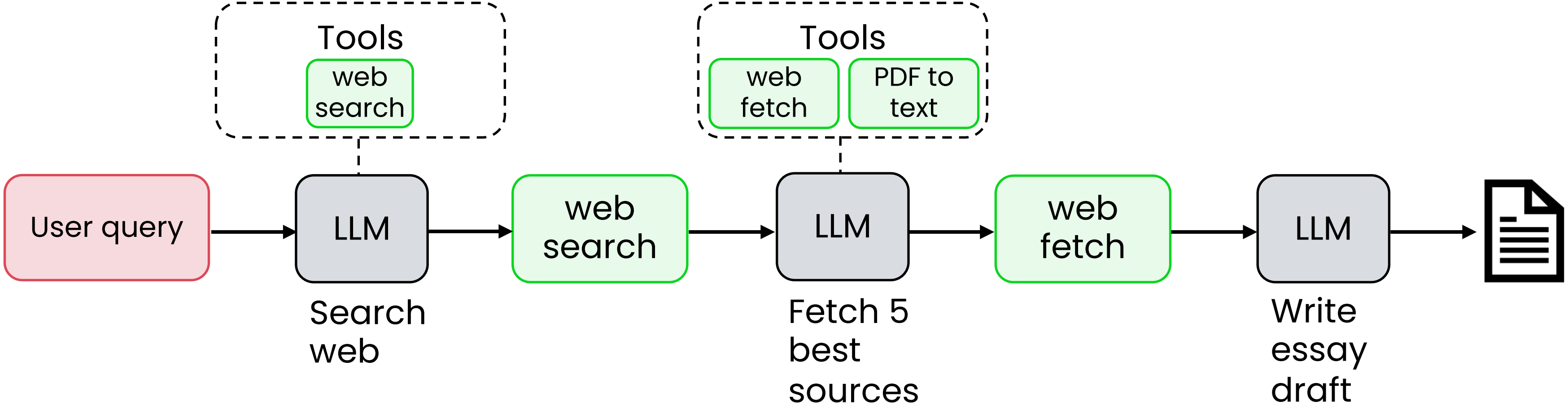
- Quick and dirty is ok to start!
- As you find places where your evals fail to capture human judgement as to what system is better, use that as an opportunity to improve the metric
- Look for places where performance is worse than humans

Practical Tips for Building Agentic AI

Error Analysis and prioritizing next steps

DeepLearning.AI

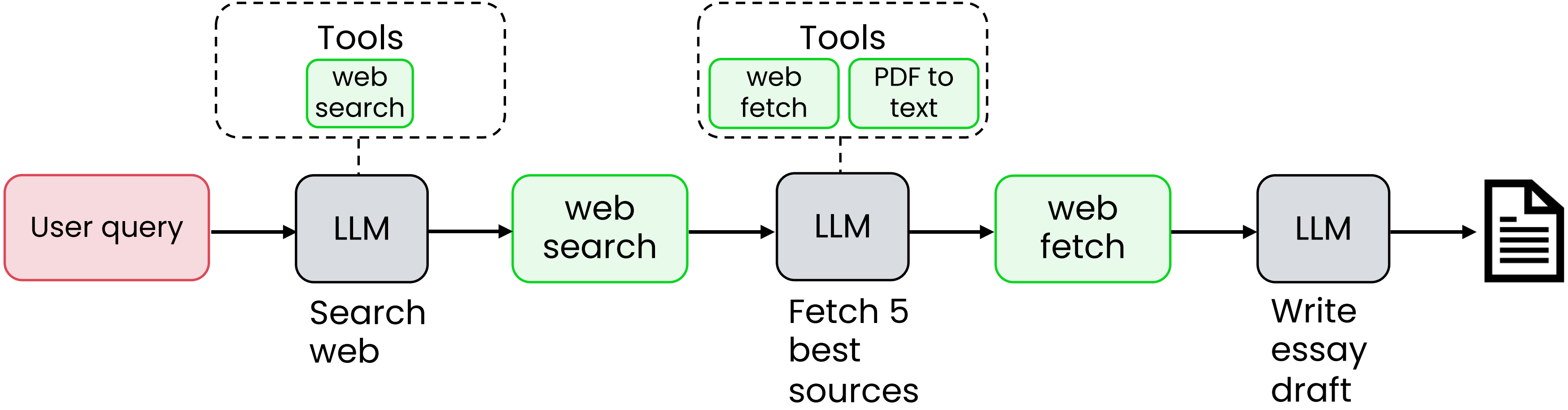
Example: research agent



Prompt	Issues
Recent black hole science	Missed high-profile result that had lots of news coverage
Renting vs buying a home in Seattle?	Seems to do a good job
Robotics for harvesting fruit	Didn't mention leading equipment company

Observed error mode: Sometimes misses key points a human would make

Example: research agent



Possible causes

Bad search terms?

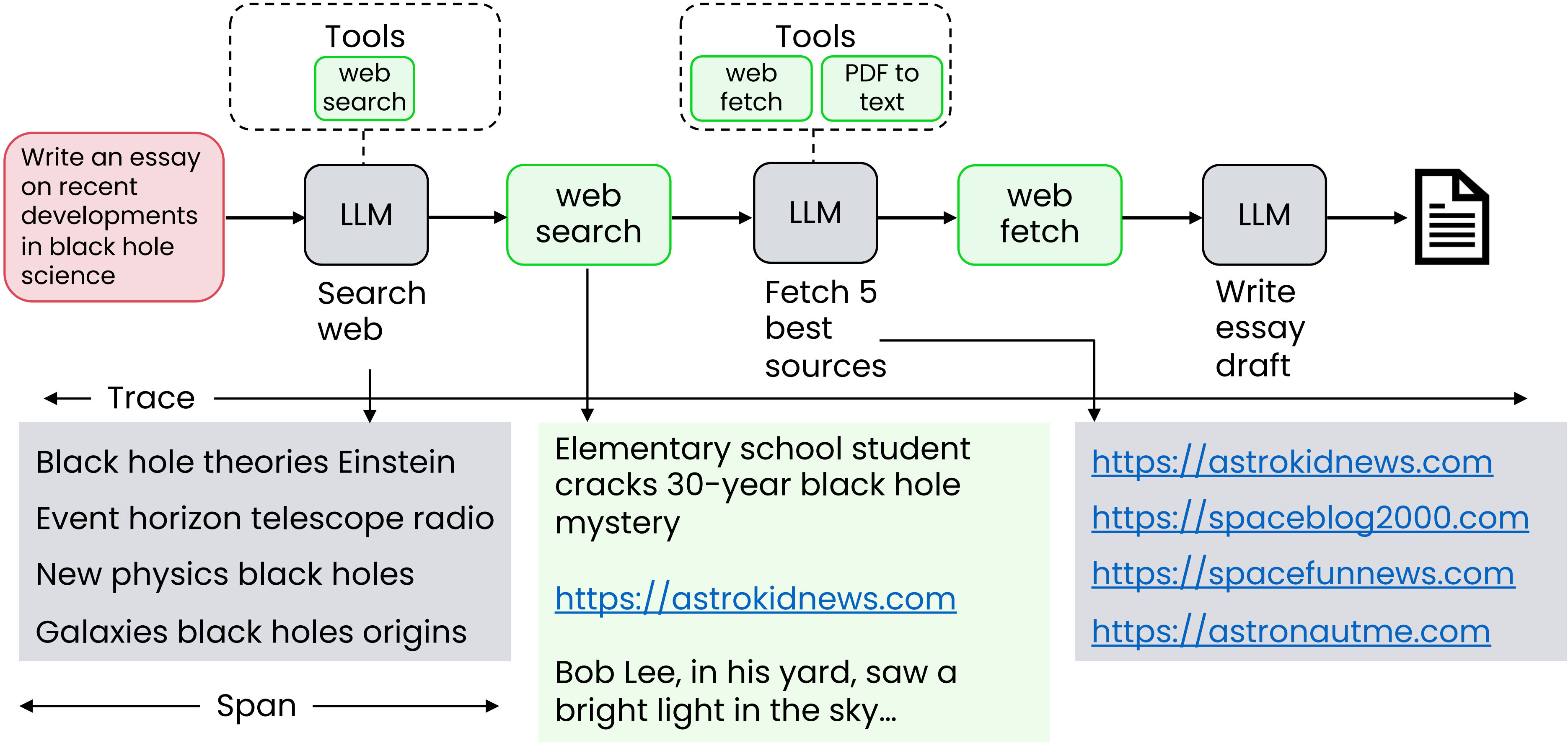
Low quality search results?

Poor selection of sources?

Bad reasoning over texts?

Examine traces to better understand each step in the workflow

Looking at traces



Write an essay on recent developments in black hole science

Tools
web search

LLM
Search web

web search

Tools
web fetch
PDF to text

LLM
Fetch 5 best sources

web fetch

LLM
Write essay draft

← Trace →

Black hole theories Einstein

Event horizon telescope radio

New physics black holes

Galaxies black holes origins

← Span →

Elementary school student cracks 30-year black hole mystery

https://astrokidnews.com

Bob Lee, in his yard, saw a bright light in the sky...

https://astrokidnews.com

https://spaceblog2000.com

https://spacefunnews.com

https://astronautme.com

Counting up the errors

Prompt	Search terms	Search results	Picking 5 best sources	…	…
Recent developments in black hole science		Too many blog posts, not enough papers			
Renting vs buying a home in Seattle			Missed well-known blog		
Robotics for harvesting fruit	Terms too generic	Website for elementary school students			
…	…	…	…		
Batteries for electric vehicles		Only selected US-based companies	Missed magazine		
	5%	**45%**	**10%**	…	…

Tips for error analysis

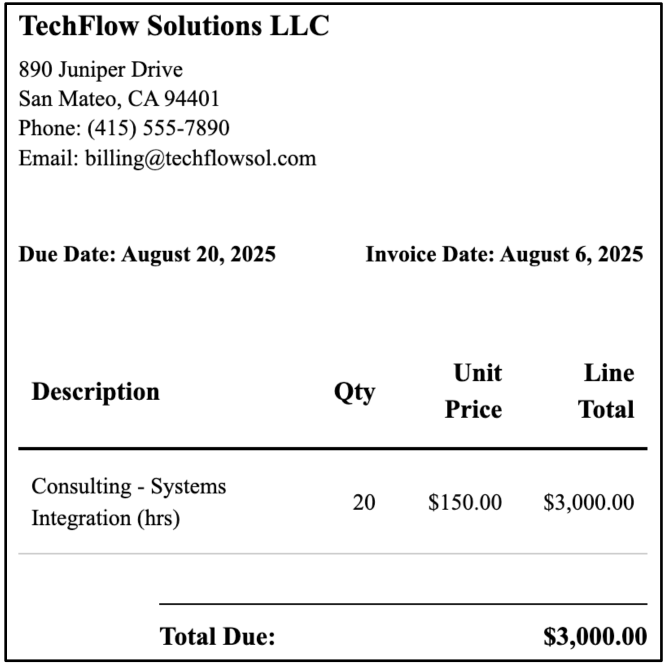
- Develop a habit of looking at traces
- Carry out error analysis to figure out what component performed poorly, leading to a poor final output
- Use error analysis output to decide where to focus efforts

Practical Tips for Building Agentic AI

More error analysis examples

DeepLearning.AI

Example: Invoice processing workflow



TechFlow Solutions LLC

890 Juniper Drive
San Mateo, CA 94401
Phone: (415) 555-7890
Email: billing@techflowsol.com

Due Date: August 20, 2025 **Invoice Date: August 6, 2025**

Description	Qty	Unit Price	Line Total
Consulting - Systems Integration (hrs)	20	$150.00	$3,000.00

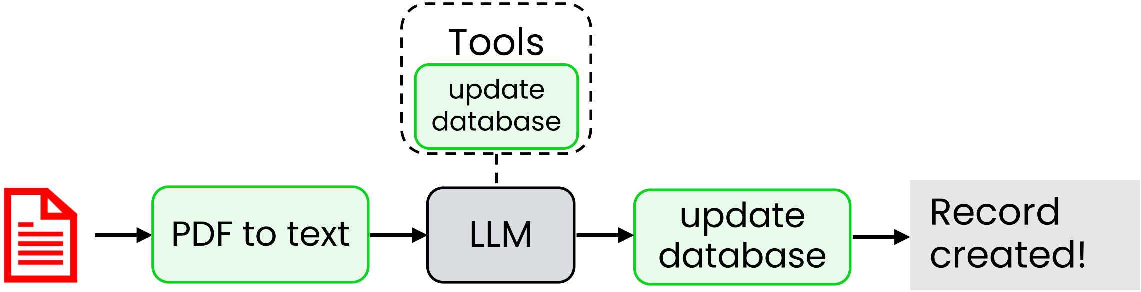
Total Due: **$3,000.00**

4 required fields:

- Biller
- Biller address
- Amount due
- Due date

Steps:

1. Identify required fields
2. Record in database



To carry out error analysis, focus on examples where performance is subpar

Counting up the errors

- Select 10-100 invoices for which the agentic workflow extracted the wrong due date

Input	PDF-to-text	LLM data extraction
Invoice 1	Errors in extraction	
Invoice 2		Wrong date selected
Invoice 3		Wrong data selected
…	…	…
Invoice 20	Errors in extraction	Wrong data selected
	15%	**87%**

Example: Responding to customer email

From: Susan Jones
Subject: Wrong item shipped

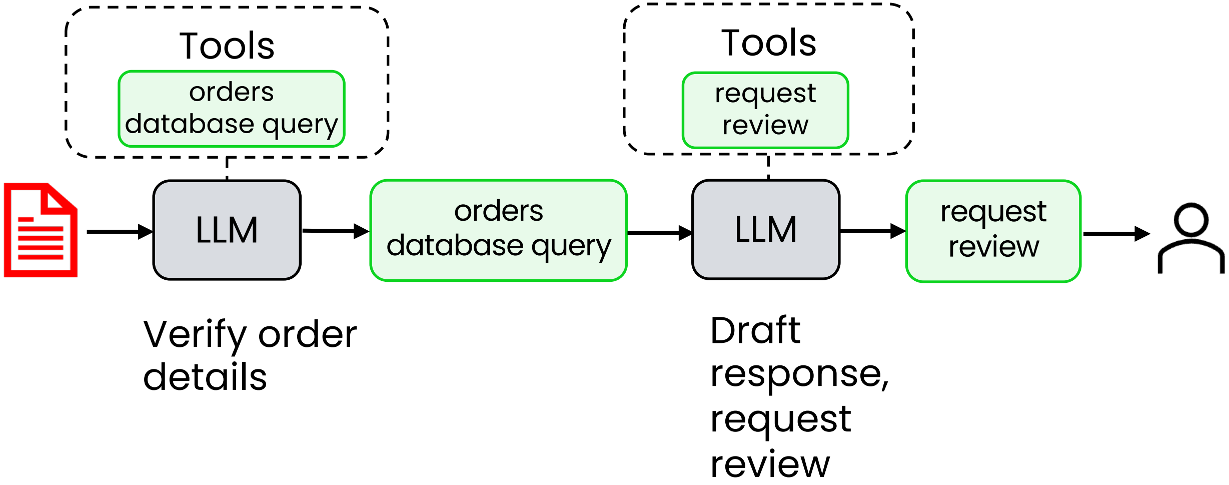
I ordered a blue KitchenPro blender (Order #8847) but received a red toaster instead.

I need the blender for my daughter's birthday party this weekend. Can you help?

Susan

Steps:

1. Extract key information
2. Find relevant customer records
3. Draft response for human review



Tools

orders database query

LLM

Verify order details

orders database query

Tools

request review

LLM

Draft response, request review

request review

Counting up the errors

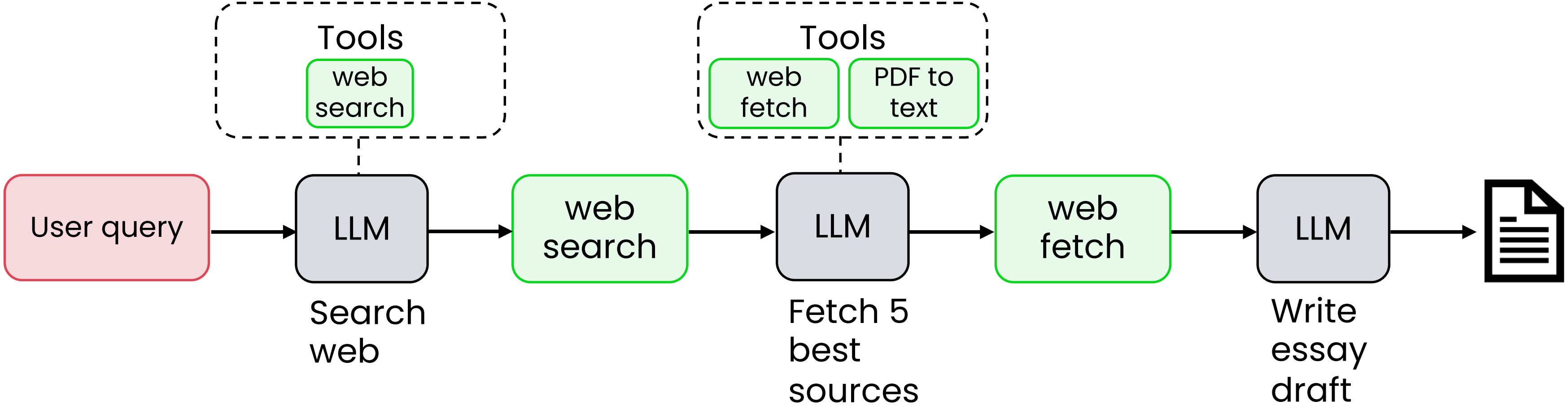
Input	LLM-drafted query	Orders database query	LLM-drafted email
Email 1	Wrong table		
Email 2		Error in database entry	Didn’t address details of order
Email 3	Incorrect math		
…	…	…	…
Email 50			Defensive tone
	75%	**4%**	**30%**

Practical Tips for Building Agentic AI

Component-level evaluations

DeepLearning.AI

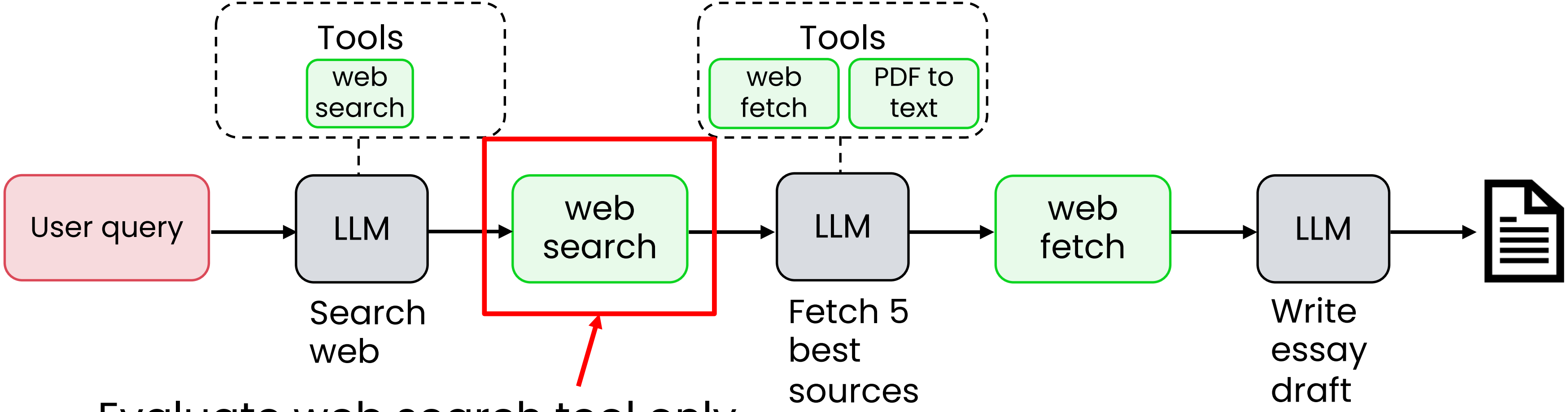
Example: research agent



Prompt	Issues
Recent black hole science	Missed high-profile result that had lots of news coverage
Renting vs buying a home in Seattle?	Seems to do a good job
Robotics for harvesting fruit	Didn't mention leading equipment company

End-to-end eval is expensive!

Example: research agent



Evaluate web search tool only

- Create a list of gold standard web resources
- Write code that calculates how many results correspond to gold standard websites e.g. F1-score
- Track as you vary hyperparameters: e.g., search engine, number of results, dates

Benefits of component-level evaluations

- Can provide clearer signal for specific errors
  - Avoid the noise in end-to-end system
- More efficient for focused team to optimize
  - Work on smaller, more targeted problems faster

Practical Tips for Building Agentic AI

How to address problems you identify

DeepLearning.AI

Improving non-LLM component performance

E.g. web search, text retrieval for RAG, code execution, trained ML model (for speech recognition, people detection, etc.)

• Tune hyperparameters of component	Web search: Number of results, date range RAG: Change similarity threshold, chunk size ML models: Detection threshold
• Replace the component	Try a different web search engine, RAG provider, etc.

Improving LLM component performance

• Improve your prompts	Add more explicit instructions. Add one or more concrete example to the prompt (few-shot prompting)
• Try a new model	Try multiple LLMs and use evals to pick the best
• Split up the step	Decompose the task into smaller steps
• Fine-tune a model	Fine tune on your internal data to improve performance

Instruction following

Summary of customer call:

On July 14, 2023, Jessica Alvarez (SSN: 555-44-3333) of 1024 Maple Ridge Lane, Boulder, CO 80301, submitted a support ticket....

Prompt

Identify all cases of personally identifiable information (PII) in the text below.

Then return a list of the identified PII classified by type, and then redact all the identified PII with "*****".

Separate the list and the redacted text with "REDACTED: ".
{text}

Instruction following results (Llama 3.1 8B)

The identified personally identifiable information (PII) classified by type is:

- Social Security Number (SSN): 524-18-7629
- Address: 1024 Maple Ridge Lane, Boulder, CO 80301

Redacted text with "*****" replacing the identified PII:

On July 14, 2023, ***** (SSN: *****), residing at *****, Boulder, CO 80301, submitted a support ticket…

List of identified PII classified by type
- SSN: 524-18-7629
- Address: 1024 Maple Ridge Lane, Boulder, CO 80301

Didn’t follow the instructions properly

Missed some PII

Instruction following results (GPT-5)

Identified PII (type → value):
1. Full Name → Jessica T. Alvarez
2. Social Security Number → 524-18-7629
3. Physical Address → 1024 Maple Ridge Lane, Boulder, CO 80301

REDACTED:
On July 14, 2023, ***** (SSN: *****), residing at *****, submitted a support ticket…

Followed formatting instruction

Identified all PII

Developing intuition for model intelligence

- Play with models often
  - Having a personal set of evals might be helpful
  - Read other people’s prompts for ideas of how to best use models

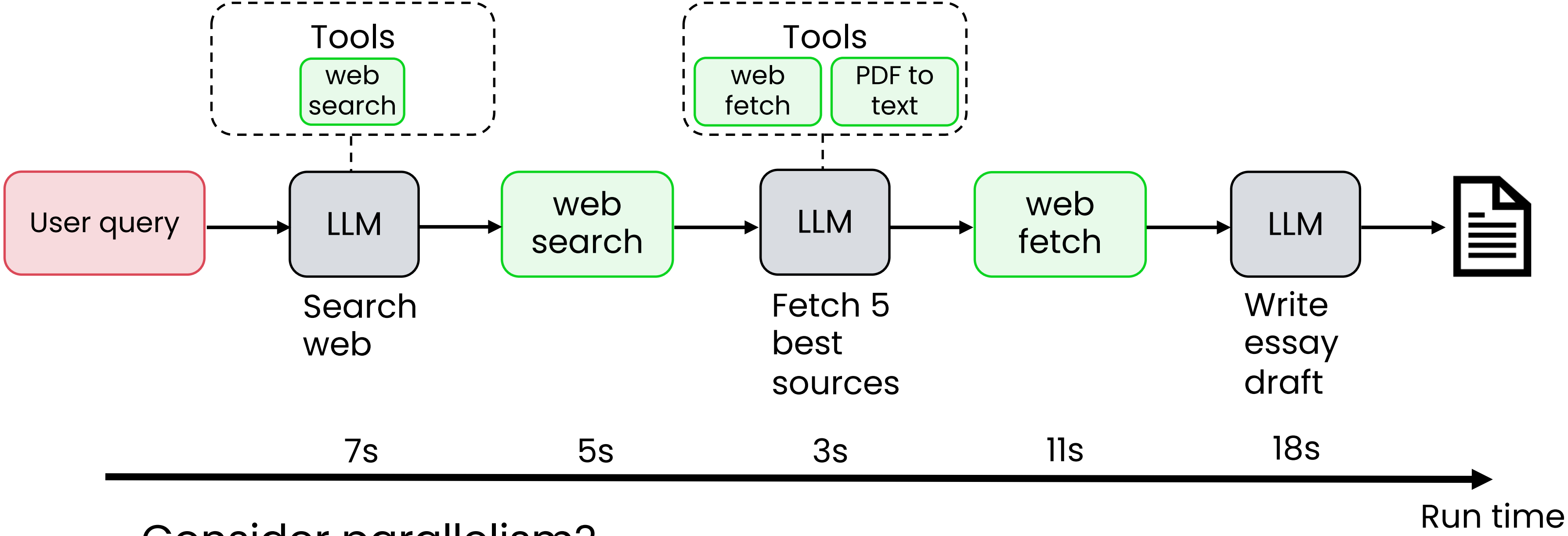
- Use different models in your agentic workflows
  - Which models work for which types of tasks?
  - aisuite makes it easy to quickly swap out models

Practical Tips for Building Agentic AI

Latency, cost optimization

DeepLearning.AI

Example: research agent



Consider parallelism?

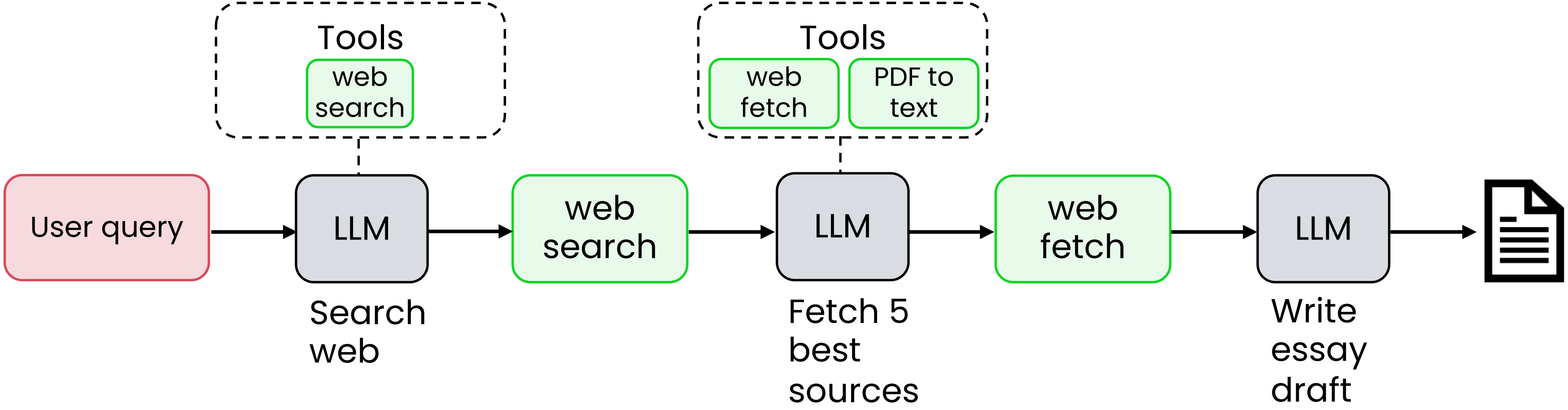
LLM steps too long?

- Try smaller/less intelligent model, or faster LLM provider

Costing your workflow

- LLM steps (pay per token)
- Any API-calling tools (pay per API call)
- Compute steps (based on server capacity/cost)

Example: research agent



Tokens:
$0.0004

API call:
$0.016

Tokens:
$0.0004

API call:
$0.002

PDF-to-text:
$0.03

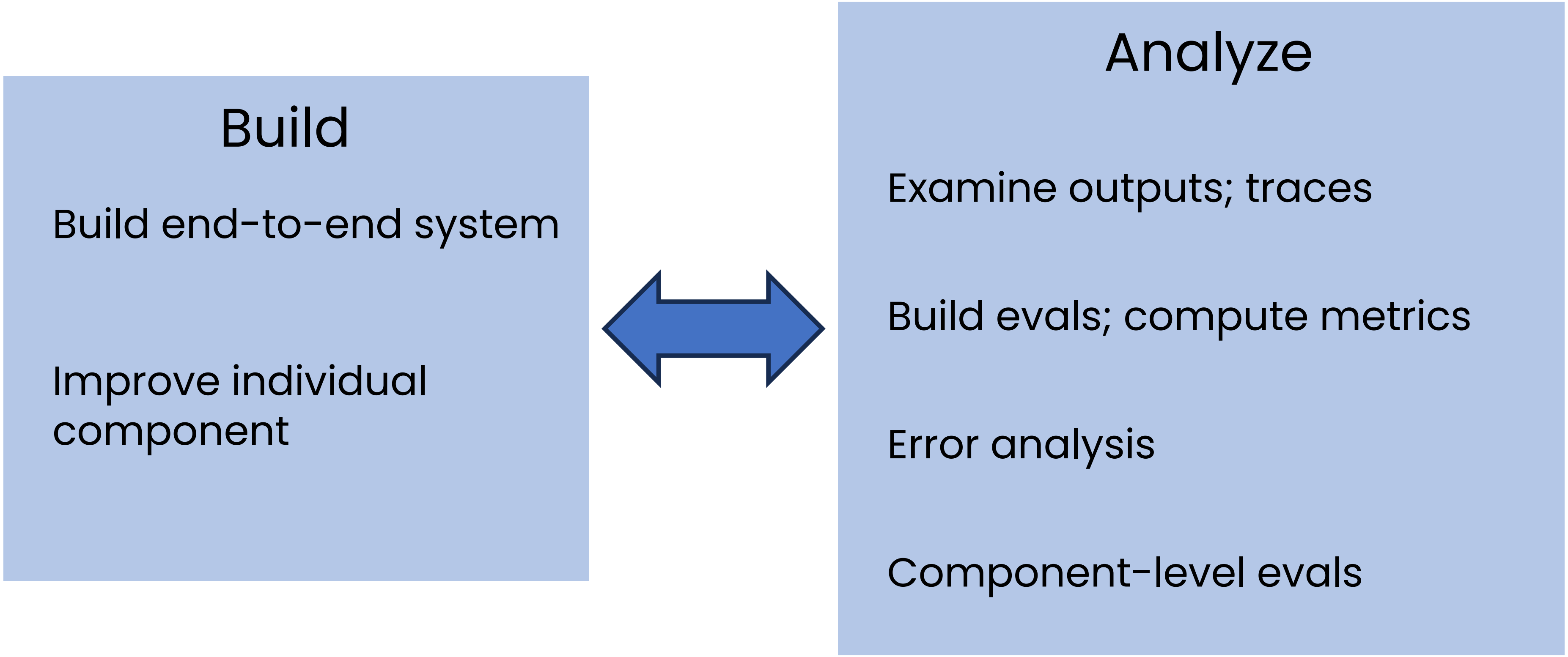
Tokens:
$0.009

Practical Tips for Building Agentic AI

Development process summary

DeepLearning.AI

Development process summary



Build

Build end-to-end system

Improve individual component

Analyze

Examine outputs; traces

Build evals; compute metrics

Error analysis

Component-level evals

End of M4

DeepLearning.AI