

Steering Semantic Data Processing with DocWrangler



A system for LLM-powered data processing

docetl.org



Shreya Shankar^{1*}, Bhavya Chopra^{1*}, Mawil Hasan¹, Stephen Lee¹
Bjoern Hartmann¹, Joseph M. Hellerstein¹, Aditya G. Parameswaran¹, Eugene Wu²
UC Berkeley¹ and Columbia University²

April 2025

DocETL: A System for Unstructured Data Processing

Launched ~6 mos ago

github.com/ucbepic/docetl

1.7k 

400+ 


No/Low-Code Interface

Declarative YAML interface and operator suite **accessible to non-programmers**

Agentic Optimizer*

Improves output accuracy and quality by intelligently and automatically **decomposing complex tasks**

We're Just Getting Started!

 Civic Engagement

 Forensic Psychiatry

 Email Analysis

 Mining Law Articles

 Summarizing educational resources

*We currently focus on optimizing accuracy, not cost.

Case Study: Scottish Climate Intelligence Services

<https://www.climateintelligenceservice.scot/climate-action-platform>

**Have: Unstructured Data
(Climate reports)**



Want: Table

The image shows a sample of unstructured climate report data. It includes a table with columns for 'Year' and 'NET reduction (%)'. The table data is as follows:

Year	NET reduction (%)
2010-11	747
2011-12	719
2012-13	691
2013-14	663
2014-15	635
2015-16	607
2016-17	579
2017-18	551
2018-19	523
2019-20	495

Other elements in the sample include text paragraphs, a line chart showing 'Total Tonnage Waste Landfilled' from 2011 to 2020, and various headers and footers from the report.

Local Authority	Intervention	KPIs	Category
Highland	E-Cargo bike pilot scheme for Council business travel	Increase in cycling	Transport
Shetland	Lower food waste	Food waste reduced by 33% by 2025	Waste
Lanarkshire	Wall insulation - Assess priority, assess feasibility, install cavity or insulation	Adoption	Energy

! Challenges
 Multiple document types
 Very long (5-200pg) & no consistent structure across them

! Challenges
 Complex reasoning required
 Cross-document analysis

DocETL: LLMs 🤝 MapReduce

docetl.org



map (for each doc)

Extract the local authority & describe any interventions (action, KPI, category)

Highland: E-Cargo bike pilot scheme for Council business travel...

Highland: Decarbonize all publicly-owned buildings...

Shetland: lower food waste...

reduce

Group by local authority; summarize and climate interventions patterns

Summary of reports under Highland Council:

E-Cargo bike scheme reducing council travel emissions while building decarbonization progressing toward 2038 zero-emission target...

Summary for Shetland Council:

Implementing comprehensive food waste reduction plan through weekly collection service, community composting sites...

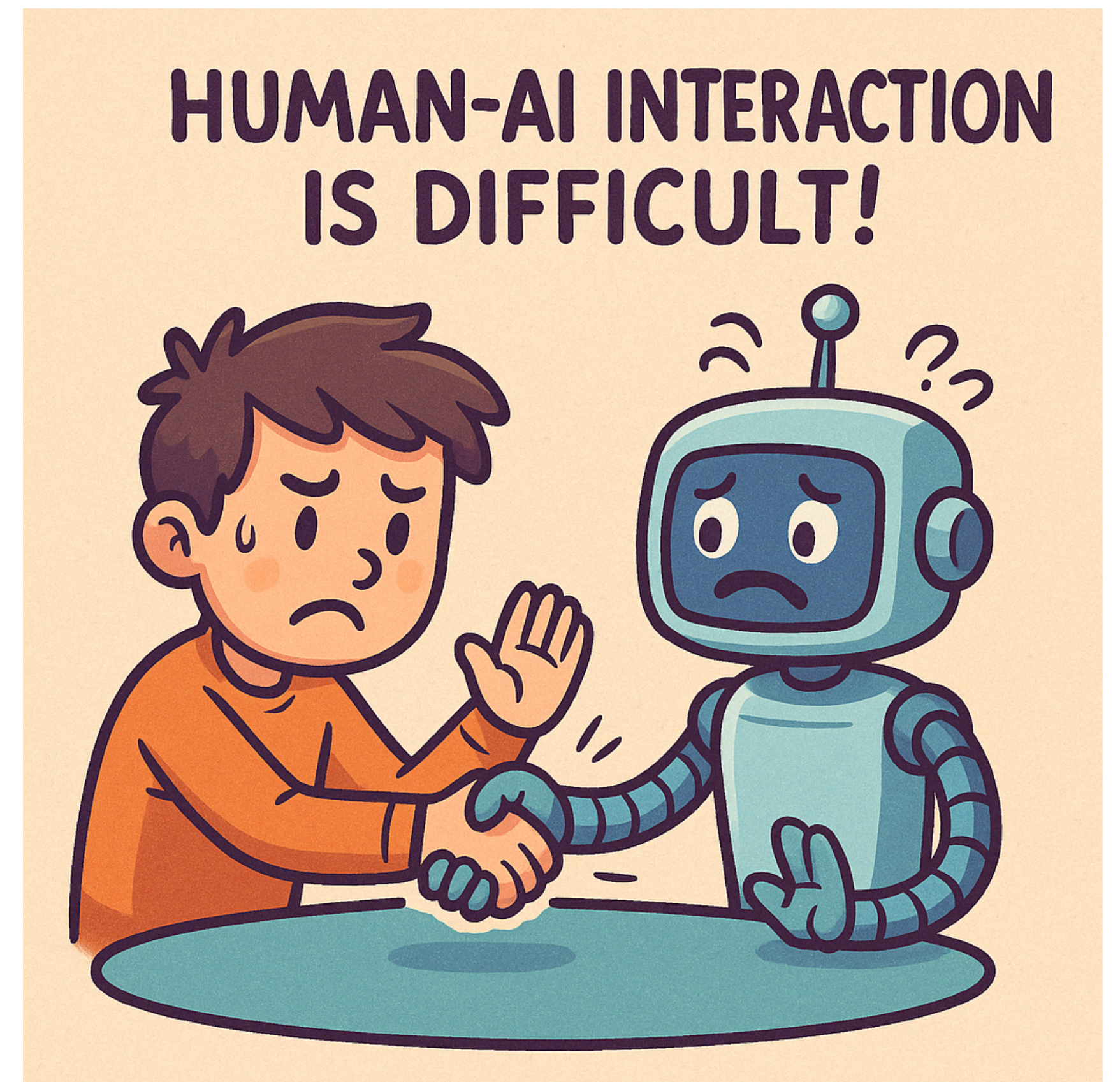
Operations defined by prompts

Operations executed by LLMs

I will call this "semantic data processing"

DocETL is great but not sufficient

- SCIS loved...
 - *“The AI impressively **took so little information fed to it** (ie just the transition elements and PDFs for a first pass)”*
 - **“managed to understand** what the other columns were to a high degree”
- ...but they had to build a **human-in-the-loop**, iterative process to arrive at the right analysis
 - *Hired a data engineer to build out DocETL pipelines*
- We are lacking an interface to author and iterate on DocETL pipelines.
- DocWrangler, an IDE for semantic data processing, is our first attempt.



DocWrangler: An IDE for Semantic Data Processing

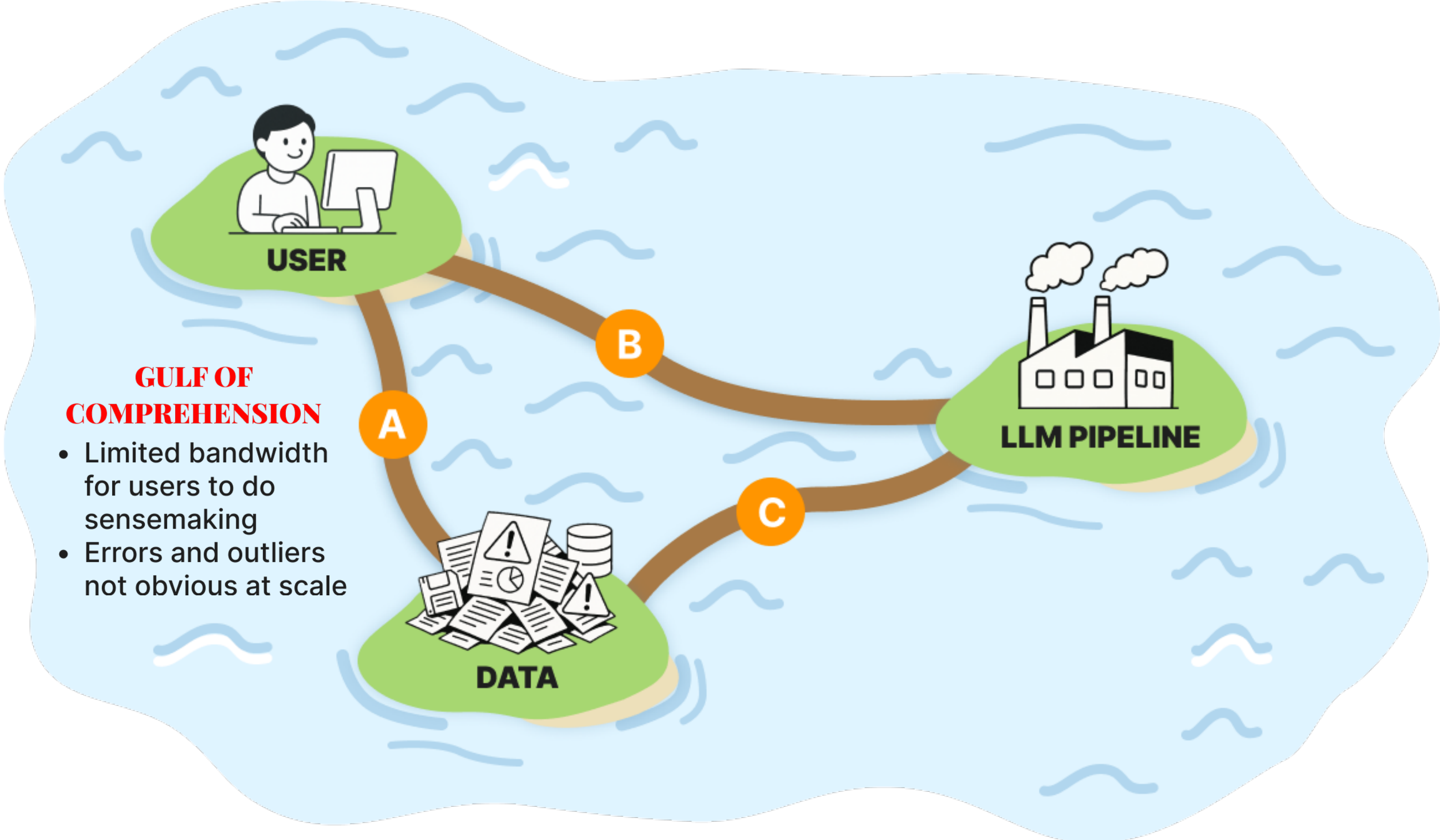
docetl.org/playground

The screenshot displays the DocWrangler IDE interface. The top navigation bar includes 'File', 'Edit', 'Help', 'Quick Save', and 'Info'. The main workspace is divided into several sections:

- FILES:** A sidebar on the left showing a file named 'medical_transcripts.json'.
- MedicalAnalysis:** The central workspace contains a pipeline editor. A prompt is visible: 'extract discomfort information from the medical transcript. identify the discomfort level (low, medium, high), provide a brief description of the discomfort, and list the symptoms the patient complains about.' Below the prompt, an 'Output Schema' is defined with fields: 'discomfort_level' (enum with values 'low', 'medium', 'high'), 'discomfort_description' (string), and 'symptoms' (list). A 'PDF URL Key' field is also present.
- OUTPUT - extract_discomfort_symptoms:** A table view at the bottom shows the results of the pipeline. It includes columns for 'symptoms', 'discomfort_level', 'discomfort_description', 'src', 'tgt', and 'file'. The first row shows an array of symptoms: 'nasal congestion', 'high blood pressure', and 'fluid in legs'. The 'discomfort_level' is 'medium'. The 'discomfort_description' is a detailed medical note. The 'src' and 'tgt' columns contain the original and processed text, respectively. The 'file' column shows the source file path.
- Dataset Statistics:** A sidebar on the right provides statistics for the 'medical_transcripts.json' dataset, including 'Documents: 87', 'Average Words: 1,626', 'Min Words: 714', 'Max Words: 3,376', and 'Std Deviation: 465'. A 'Word Count Distribution' histogram is also shown.

- A: Pipeline editor
- B: Input & Output inspector

Interaction “Gulfs” in Semantic Data Processing



Illustrating the Gulf of Comprehension

“Extract climate interventions (category, desc, KPI) from the document”

LLM Output

“Category Transport: Installing 24 EV charge points to encourage the uptake of EVs across North Ayrshire.”



Human Refinement

“Hm maybe **Transport/shift to EV should be its own category**, since many towns are doing this.”

LLM Output

“Category Transport/EV: Installing 24 EV charge points to encourage the uptake of EVs across North Ayrshire.”



Human Refinement

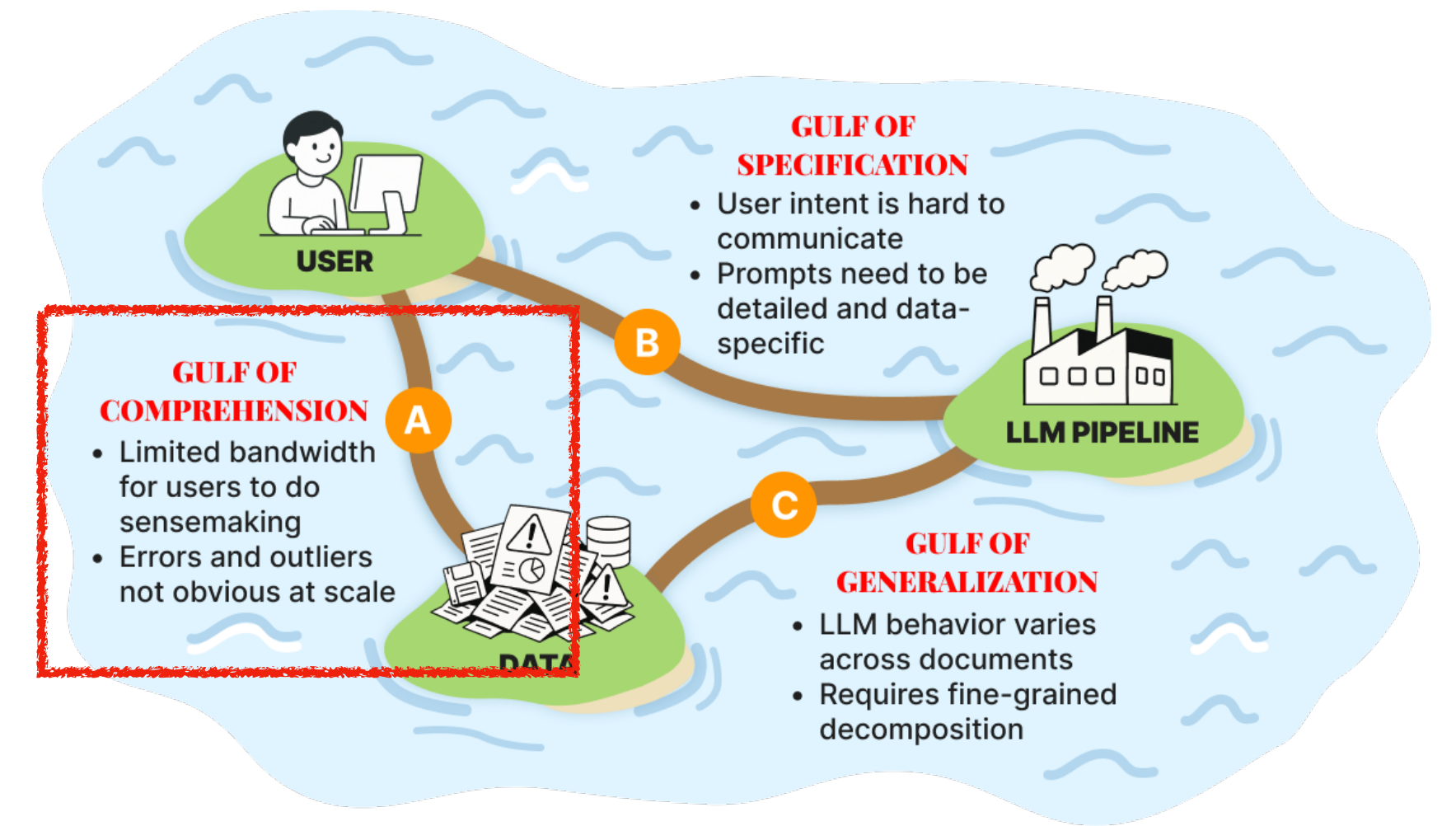
“Also explicitly indicate if any efficiency metrics (CO2 reduction) are **N/A** in the document.”

KEY ISSUE

Users need to make sense of the data to write the correct prompt.

Solution: In-Situ User Notes

Bridging the Comprehension Gulf



- DocETL user challenges
 - Keep track of notes on outputs, documents, and LLM behavior
 - Find patterns in data and behavior

KEY IDEA

We capture user notes in-situ, as they are looking at data.

Screenshots: In-Situ User Notes

A Table Visualize Input Distribution

discomfort_description

3 words avg: 18 76 words 628 words

Search in cell...

Patient has elevated blood pressure and some fluid in legs, along with a heart murmur and low pumping function of the heart.

B discomfort_description

80th percentile

Type: string-words, Distinct Values: 10 / 10, Current: 22 words, Range: 3 - 76

Search in cell...

Patient has elevated blood pressure and some fluid in legs, along with a heart murmur and low pumping function of the heart.

C Other Keys

discomfort_level

medium

file

src

[doctor] hi , martha . how are you ?

[doctor] martha is a 50-year-old fe

it's a sea... how are you doing?

D Add Notes

Your notes will help improve prompts via the Improve Prompt feature in operation settings

the discomfort level should be about how comfortable, behavioral, the patient is (not about the physical symptoms per se).

Category Add Note

E NOTES

Tip: Click in any output column to leave notes on outputs

Note: Notes are only used when clicking Improve Prompt, not in operation prompts

Search...

the discomfort level should be about how comfortable, behavioral, the patient is (not about the physical symptoms per se).

Column: discomfort_description - Row 0

Row Context

```
{
  "src": "[doctor] and why is she her",
  "tgt": "CHIEF COMPLAINT: Annual v",
  "file": "B2M07-vitalist",
  "discomfort_level": "medium",
  "symptoms": [
    "depression",
    "chronic back pain",
    "numbing in legs",
    "stiffness"
  ],
  "discomfort_description": "The pati"
}
```

Delete

I think it's just a lot has been...

OUTPUT - extract_disc

Show/Hide Columns

symptoms

Filter...

2 rows avg: 4 6 items

Search in cell...

Array (3 items)

0: "nausea"

Search in cell...

Array (5 items)

0: "knee pain"

Search in cell...

Array (6 items)

0: "back pain"

Search in cell...

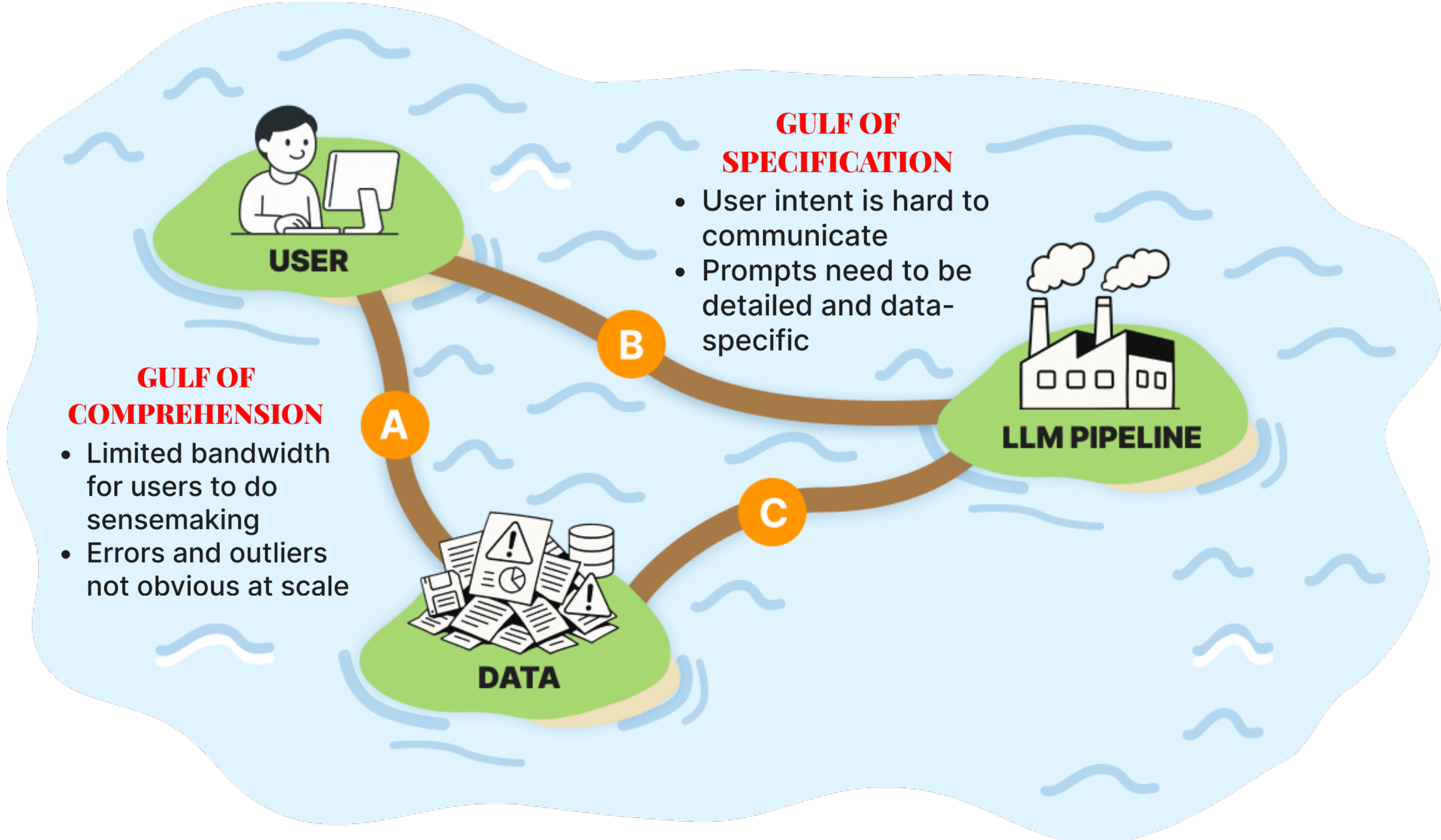
Array (3 items)

0: "back pain"

Search in cell...

Array (2 items)

Interaction Challenges in Semantic Data Processing



Illustrating the Gulf of Specification

“Extract climate interventions from the following document”

LLM Output

"North Ayrshire Council invested nearly £1m in energy efficiency measures such as LED lighting, boiler..."



Human Refinement

"Please specify this as an intervention **with category** (energy efficiency) and measurable outcome."

LLM Output

"The Council has retrofitted 14,378 street lights with energy efficient LED fittings across North Ayrshire."



Human Refinement

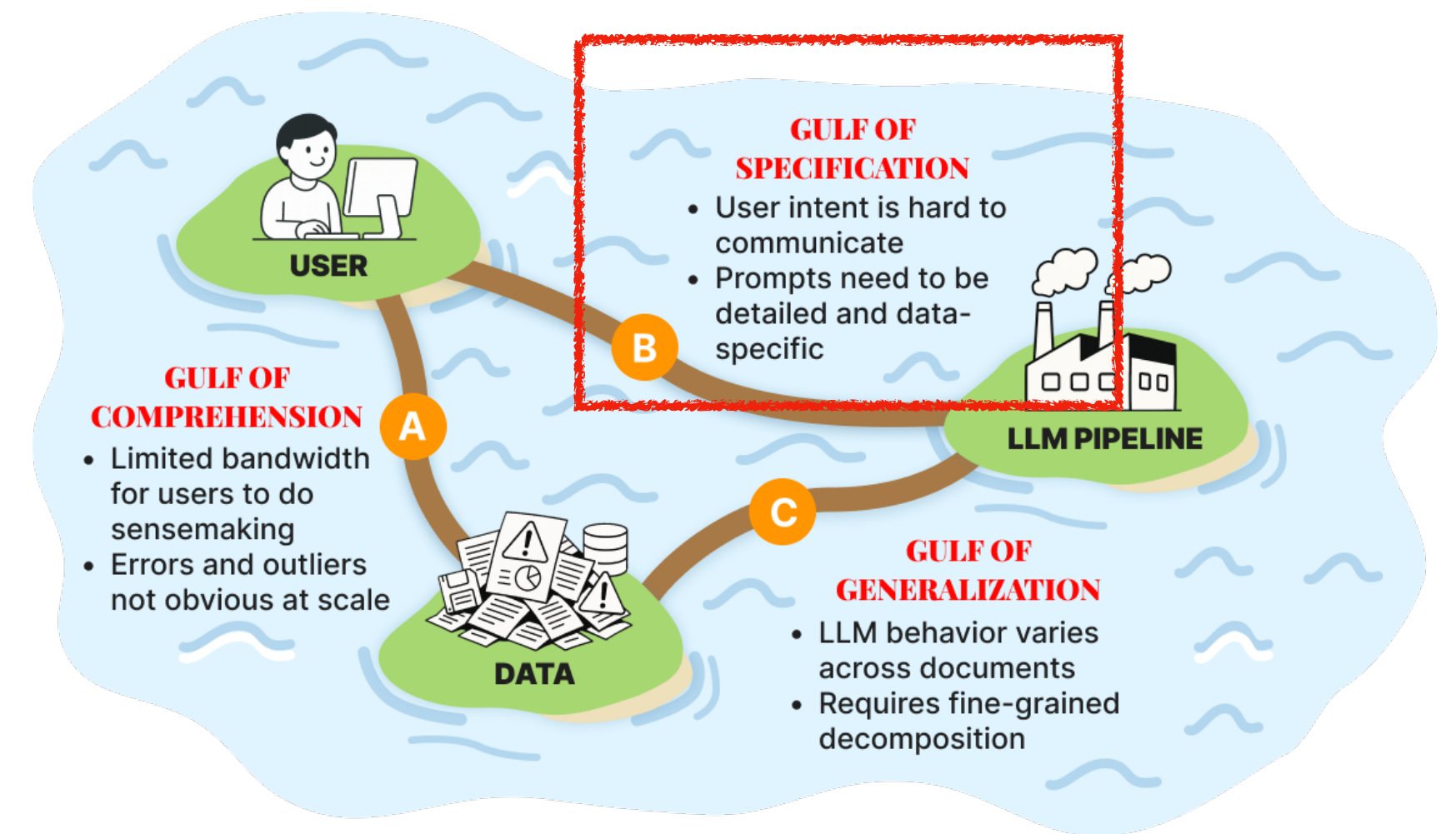
"Add the climate **impact metric** - what percentage carbon reduction was achieved through this intervention?"

KEY ISSUE

Prompts requires **thorough specification** ("climate intervention" = category, description, KPI, etc.)

Solution: Prompt Refinement

Bridging the Specification Gulf



- DocETL user challenges
 - Translate observations into actual prompt modifications
 - Adding specificity; maybe even changing the task altogether

KEY IDEA

We provide a conversational interface for AI-assisted prompt edits.

Screenshots: Assisted Prompt Refinement

MedicalAnalysis | Add Operation + | Stop | Run Fresh | Run

map ? gemini/gemini-2.0-flash extract_discomfort_symptoms Show Outputs Improve Prompt

extract discomfort information from the medical transcript.
{{ input.src }}

A

Improve Prompt

Select the operation you want to improve the prompt for

extract_discomfort_symptoms

Current Prompt:

extract discomfort information from the medical transcript.
{{ input.src }}

identify the discomfort level (low, medium, high), provide a brief description of the discomfort about.

Your Notes:

- the discomfort level should be about how comfortable, behaviorally, the patient is (not about the physical symptoms)
- "I think it's just a lot has been weighing on me lately" this is an example of low discomfort b/c the patient feels comfortable

Additional Instructions (optional)

Add specific instructions for improving the prompt (e.g., 'Make it more concise', 'Add more examples')

Leave blank to let the AI follow default improvement guidelines

Continue to Analysis

B

← Improve Prompt

DocWrangler is analyzing and suggesting improvements

Your current prompt provides a basic structure but lacks specificity and clarity regarding the discomfort level's behavioral aspect. The feedback you've received emphasizes the need to focus on the patient's comfort level rather than just physical symptoms. Here's an improved version of your prompt that incorporates your feedback and adheres to best practices: This revised prompt is specific, breaks down the task into clear steps, and provides examples to guide the output. It also emphasizes the behavioral aspect of discomfort, aligning with the feedback you've received.

Prompt Changes:

extract discomfort information from the medical transcript.
Extract @discomfort information from the medical transcript.

{{ input.src }}

identify the discomfort level (low, medium, high), provide a brief description of the discomfort, and list the symptoms the patient complains about. Follow these steps:

- Identify the discomfort level: Determine if the patient's discomfort level is low, medium, or high based on their behavioral comfort rather than just physical symptoms. For example:
 - Low discomfort: "I think it's just a lot has been weighing on me lately" (indicates comfort in sharing personal details).
 - Medium discomfort: "I feel a bit stressed but managing" (indicates some discomfort but still functional).
 - High discomfort: "I'm in a lot of pain and can't focus" (indicates significant discomfort affecting functionality).
- Provide a brief description of the discomfort: Summarize the patient's discomfort in one to two sentences, focusing on their emotional and behavioral state.
- List the symptoms: Identify and list the specific symptoms the patient complains about, formatted as a bullet list.

Output format:

- Discomfort Level: [low/medium/high]
- Discomfort Description: [Brief description]
- Symptoms:
 - [Symptom 1]
 - [Symptom 2]
 - [Symptom 3]

Ensure that the discomfort level reflects the patient's behavioral comfort and not just physical symptoms.

Directly edit | Add feedback | Save and Overwrite

C

D **E** **F**

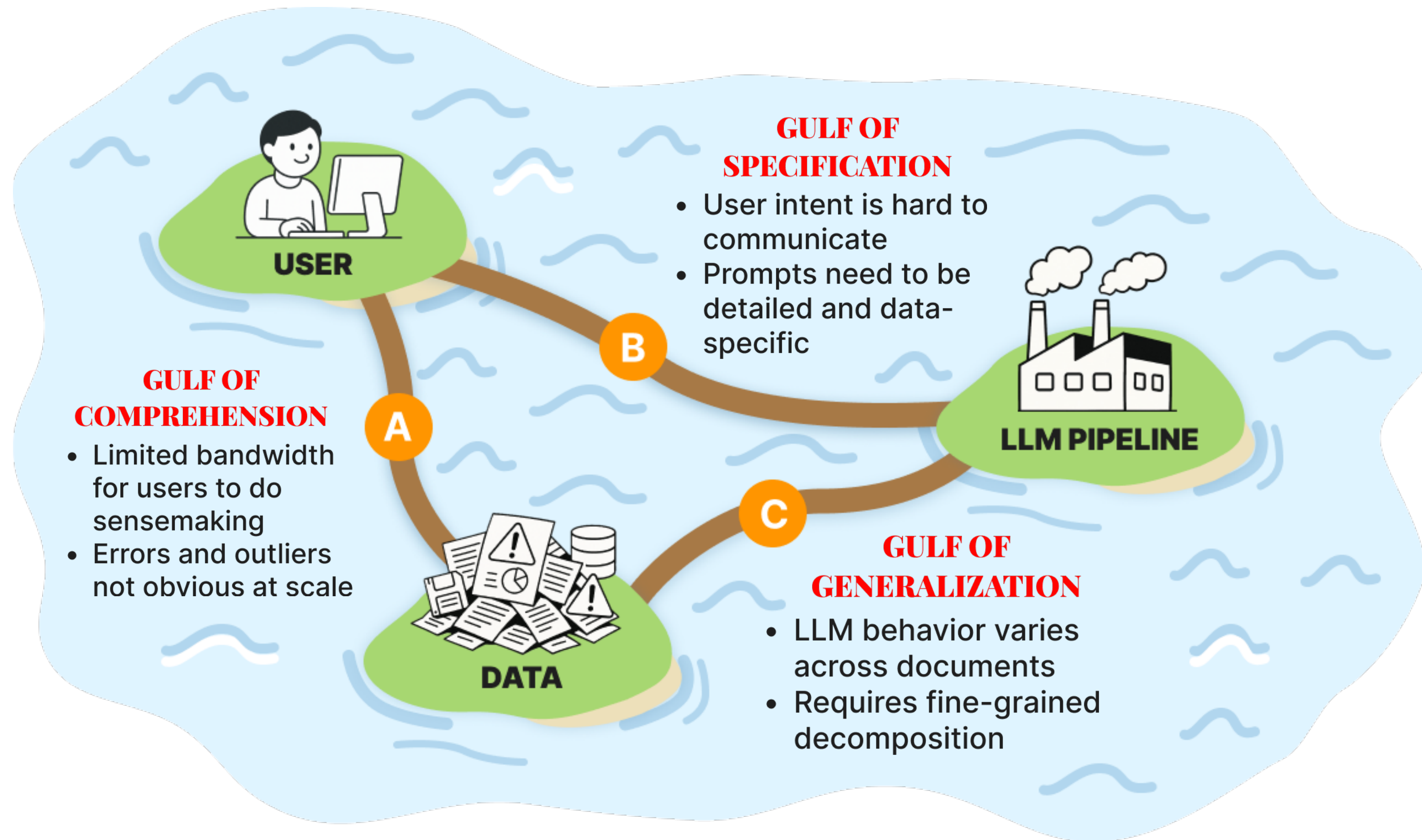
OUTPUT - extract_discomfort_symptoms | Console | Table | Visualize Input Distribution

Show/Hide Columns | Reset Widths

symptoms	discomfort_level	discomfort_description
nasal congestion	low	The patient exhibits a low level of discomfort, openly discussing her medical history, lifestyle, and

G

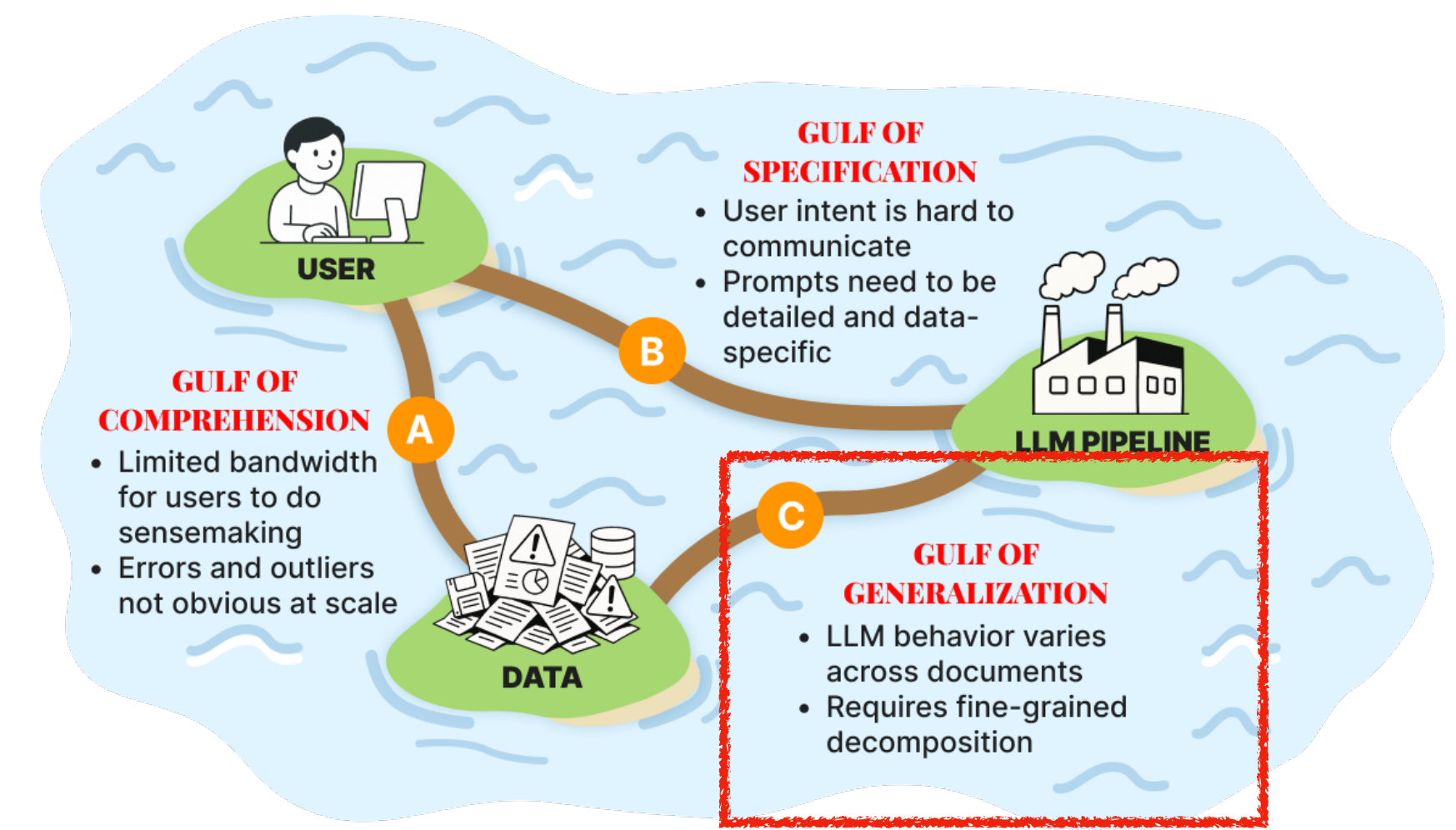
Interaction Challenges in Semantic Data Processing



Solution: Assisted Decomposition

Bridging the Generalization Gulf

- DocETL user challenges
 - Knowing when an operation is “too complex” as-is
 - Knowing how to decompose operations



KEY IDEA

We “notify” users when we think operations are too complex & why.

Screenshots: Operation Decomposition

*It might take 5+ minutes to run interactively

A Hey! Consider decomposing `extract_discomfort_symptoms`. [Click here to see why.](#)

B **Operation Too Complex**
This operation might be too complex for the LLM to handle efficiently. We recommend breaking it down into smaller, more manageable steps.

okay . how are you ?		1b9ce/896t55	"high blood pressure", "fluid in legs", "heart murmur", "slightly high cholesterol"	medical history, lifestyle, and adherence to treatment plans. She willingly answers questions about her health conditions, emotional well-being, and medication adherence, indicating a comfortable and
[doctor] i'm doing	HISTORY OF PRESENT ILLNESS			
okay . so , i know the nurse told you about				
dax . i'd like to tell dax a little bit about you , okay ?	Martha Collins is a 50-year-old female with a past medical history significant for congestive heart failure depression			
[patient] okay .				
[doctor] martha is a 50-year-old female				

Row 1 of 87

Suggested Improvements

Row 1 and Row 2 outputs correctly identify the discomfort level as low and provide clear rationales based on the patient's openness and comfort in discussing their health. However, Row 3 incorrectly identifies the discomfort level as high despite the patient being able to communicate effectively which contradicts the criteria for high discomfort. The rationale provided does not align with the patient's ability to articulate their concerns and

C **Automatically Decompose**

D **OUTPUT - extract_discomfort_symptoms**

Checking if optimization is needed... 25%

OPTIMIZING BECAUSE

In Row 1, the discomfort level is classified as low, which is appropriate given the patient's open communication and lack of evasive responses. The rationale provided is clear and aligns with the criteria for low discomfort. However, the symptoms listed include 'Vitamin D deficiency' and 'High cholesterol,' which are not explicitly mentioned as complaints by the patient in the transcript, indicating a potential oversight in symptom identification. In Row 2, the discomfort level is classified as high, which is justified based on the patient's expressions of pain and frustration regarding his treatment. The rationale is well-articulated, capturing the essence of the patient's emotional state. The symptoms listed accurately reflect the patient's complaints, although the phrase 'ow' could be considered informal and might not fit the clinical context. In Row 3, the discomfort level is classified as low, which is appropriate as the patient engages openly and does not exhibit signs of distress. The rationale is clearly explained. However, the symptoms listed do not include the patient's mention of 'nagging feeling' in the abdomen, which could be considered a significant symptom related to his discomfort. In Row 1, ensure that all symptoms mentioned by the patient are included in the output. For example, 'Vitamin D deficiency' and 'High cholesterol' should be omitted unless they are explicitly stated as concerns by the patient. Instead, focus on symptoms that the patient directly expresses discomfort about, such as 'weight gain' and 'acne.' In Row 3, include the 'nagging feeling' symptom in the output to provide a more comprehensive view of the patient's discomfort. This would enhance the accuracy of the symptom list and better reflect the patient's experience.

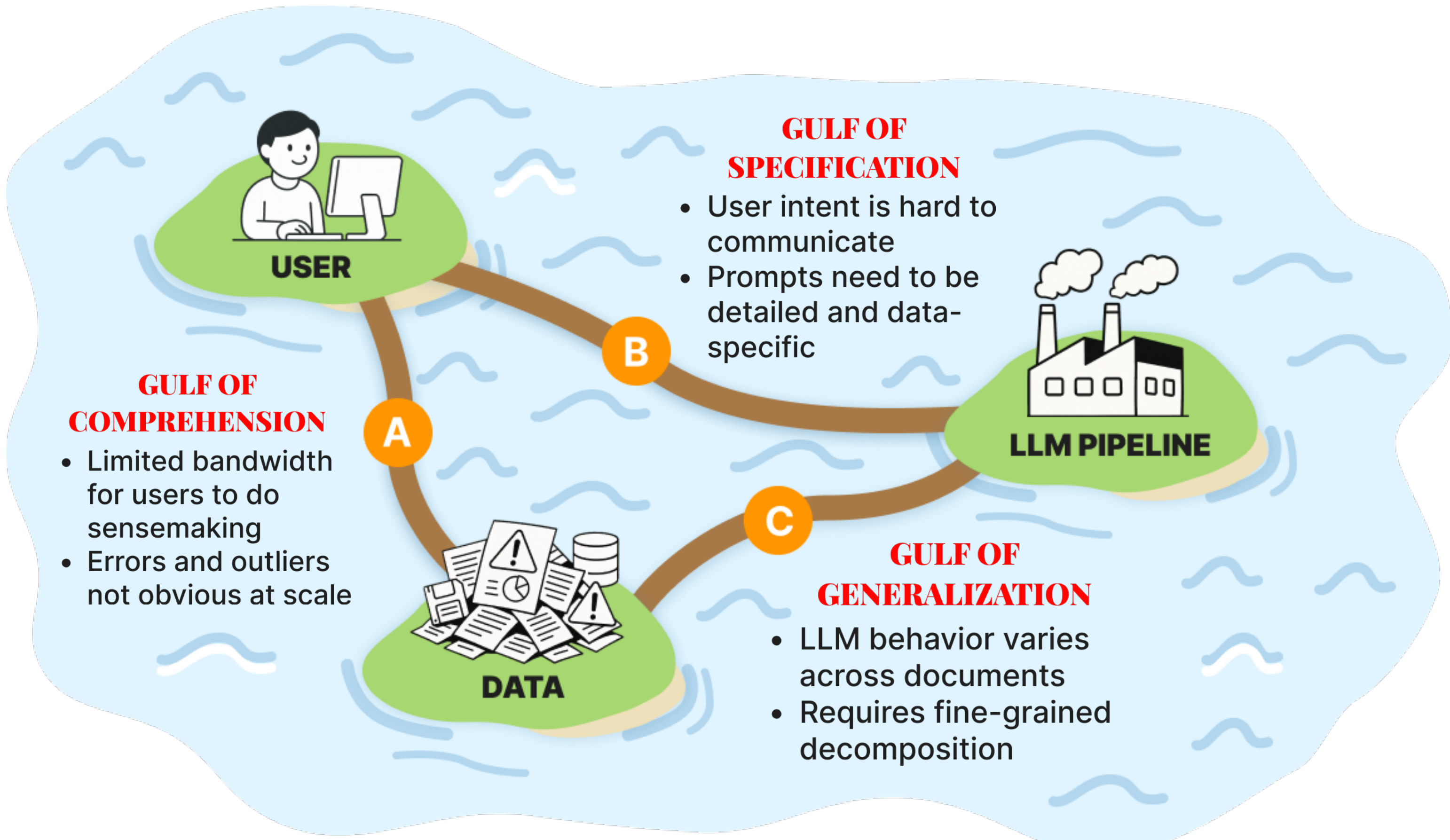
USING THIS PROMPT TO EVALUATE THE BEST PLAN

1. Does the output accurately identify the patient's social discomfort level (low, medium, or high) based on the provided criteria, and is the rationale for this classification clearly explained?
2. Are all symptoms mentioned by the patient in the transcript correctly identified and listed in the output, and is the format consistent with the specified output structure?

E **Pipeline Flow** 4 operations

- split_extract_discomfort_symptoms (split)
- gather_src_extract_discomfort_symptoms (gather)
- submap_extract_discomfort_symptoms (map)
- subreduce_extract_discomfort_symptoms (reduce)

Interaction Challenges in Semantic Data Processing

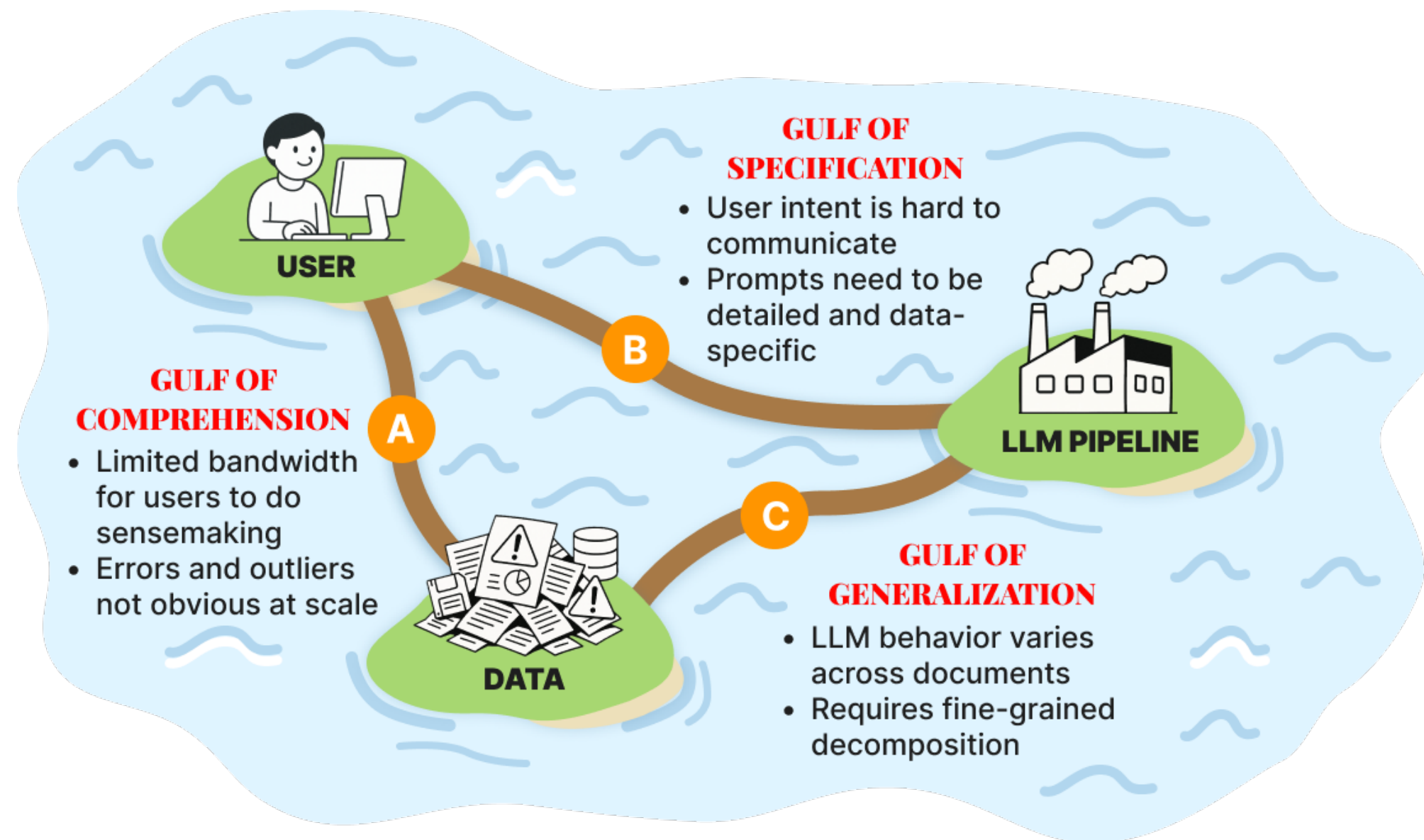


The right side of the image shows three UI mockups illustrating interaction challenges, labeled A, B, and C.

- A) Docs & Outputs:** A dashboard showing data visualizations (bar charts) for 'condition', 'discomfort_level', and 'symptoms'. An 'In-Situ User Notes' box is overlaid on the 'discomfort_level' chart, with an 'Add Note' button.
- B) Prompt Refinement:** A section titled 'Suggested Modifications for Prompt' with a list of items and buttons for 'Edit', 'Add Feedback', and 'Accept Changes'.
- C) Operation Decomposition:** A section titled 'Operation Too Complex!' with the text 'Consider decomposing extract_discomfort because...' and a 'Decompose' button.

Lessons Learned From User Studies

Participants (n=10) found DocWrangler useful and usable (80% rated 6–7 on 7pt scale)

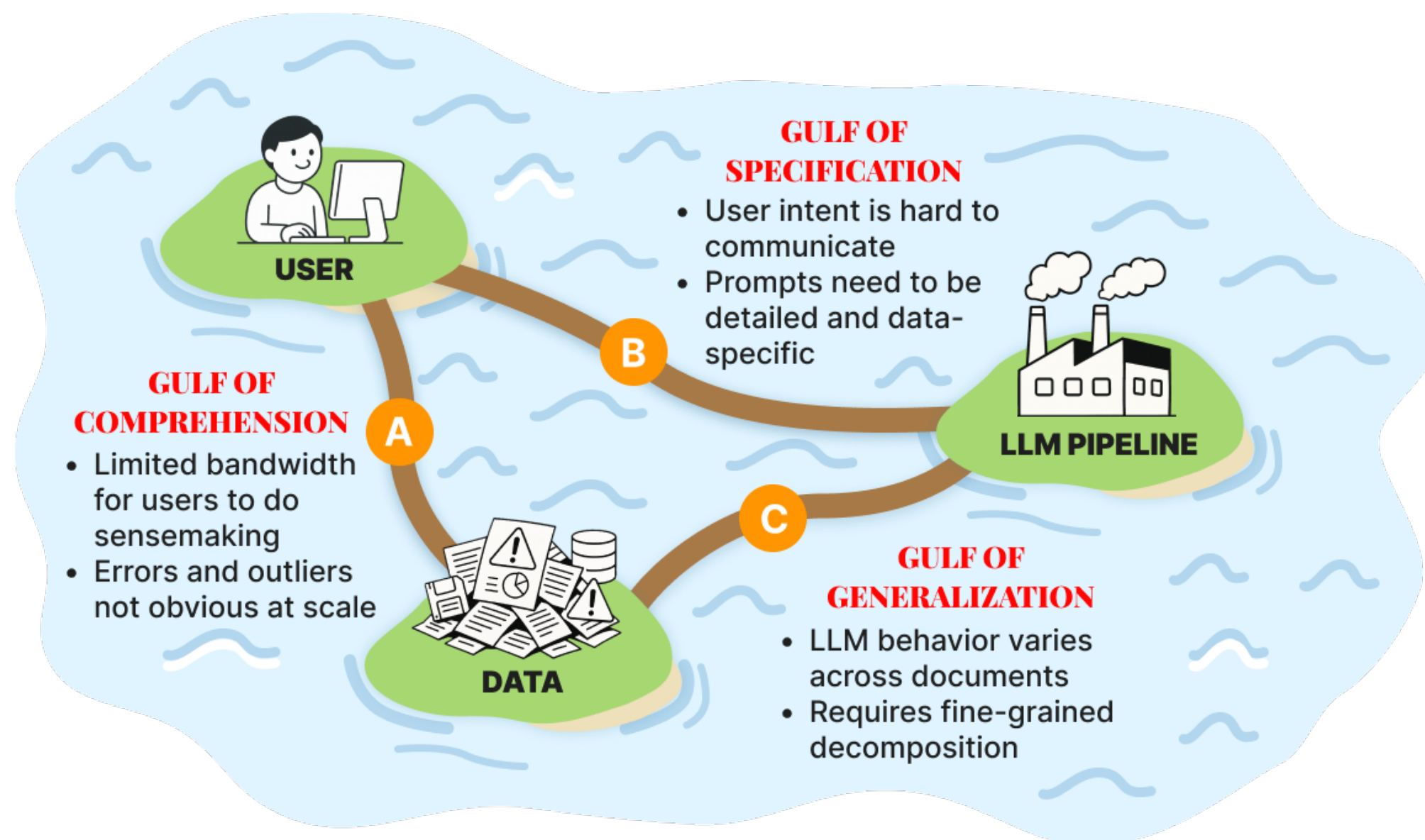


Common strategies

- Add “reasoning” or “explanation” attributes to output schemas
- Add structured attributes to open-ended operations to aid validation
- “Prompt rubber ducking” (P9) helped users clarify *what* to ask

Online Deployment Insights

1500+ pipeline runs across many domains (e.g., legal, medical, finance, education) and 9+ languages



- 90% of pipelines were “shallow” (≤ 3 ops), but some had 15–30+ ops
- 18% of pipelines used multiple models
- Pipelines evolved in 3 main ways
 - 53% more complex (e.g., more operations, model upgrades)
 - 18% actually simpler (e.g., reduced operations; open-ended prompts)
 - 29% same structure with prompt/schema changes

Takeaways

- Users need rich interfaces for *semantic data processing*
- DocWrangler bridges 3 interaction “gulfs”
 - *Comprehension*: In-situ user notes
 - *Specification*: Assisted prompt refinements
 - *Generalization*: Assisted operator decomposition
- Next: better document visualization and provenance, steering optimization

Thank
you!

shreyashankar@berkeley.edu

sh-reya.com/docwrangler-techreport.pdf