Steering Semantic Data Processing with DocWrangler

docetl.org

EPIC DATA lab
UC Berkeley

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Shreya Shankar[1*], Bhavya Chopra[1*], Mawil Hasan[1], Stephen Lee[1]
Bjoern Hartmann[1], Joseph M. Hellerstein[1], Aditya G. Parameswaran[1], Eugene Wu[2]
UC Berkeley[1] and Columbia University[2]
April 2025

DocETL: A System for Unstructured Data Processing

Launched ~6 mos ago | github.com/ucbepic/docetl | 1.7k ⭐ | 400+ Discord

🧩 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 currently focus on optimizing accuracy, not cost.

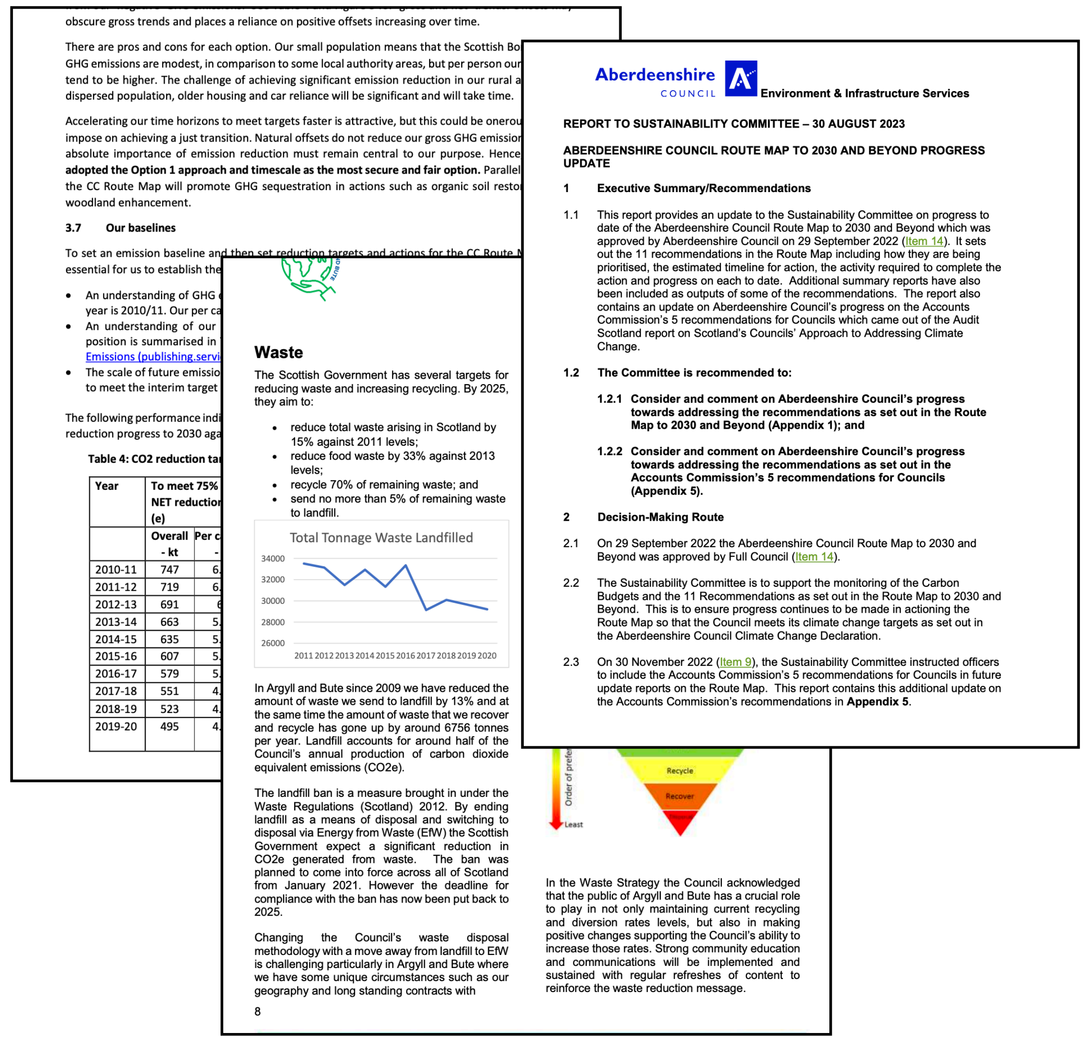
We're Just Getting Started! 🚀

- 🏛️ Civic Engagement
- 🔍 Forensic Psychiatry
- 📧 Email Analysis
- ⚖️ Mining Law Articles
- 📚 Summarizing educational resources

Case Study: Scottish Climate Intelligence Services

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

Have: Unstructured Data (Climate reports) → Want: Table



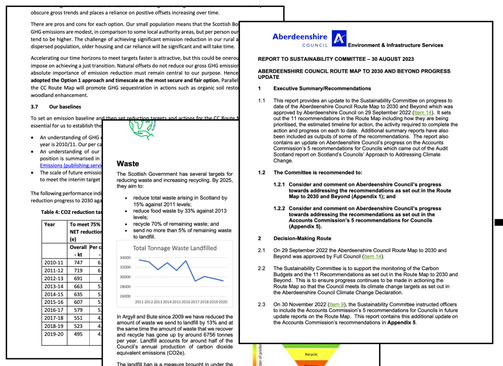
Local Authority	Intervention	KPIs	Category
Highland	E-Cargo bike pilot scheme for Council business travel	Increase in cycling	Transport
Shetland	Lower food waste	Food waste reduced by 33% by 2025	Waste
Lanarkshire	Wall insulation - Assess priority, assess feasibility, install cavity or insultation	Adoption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

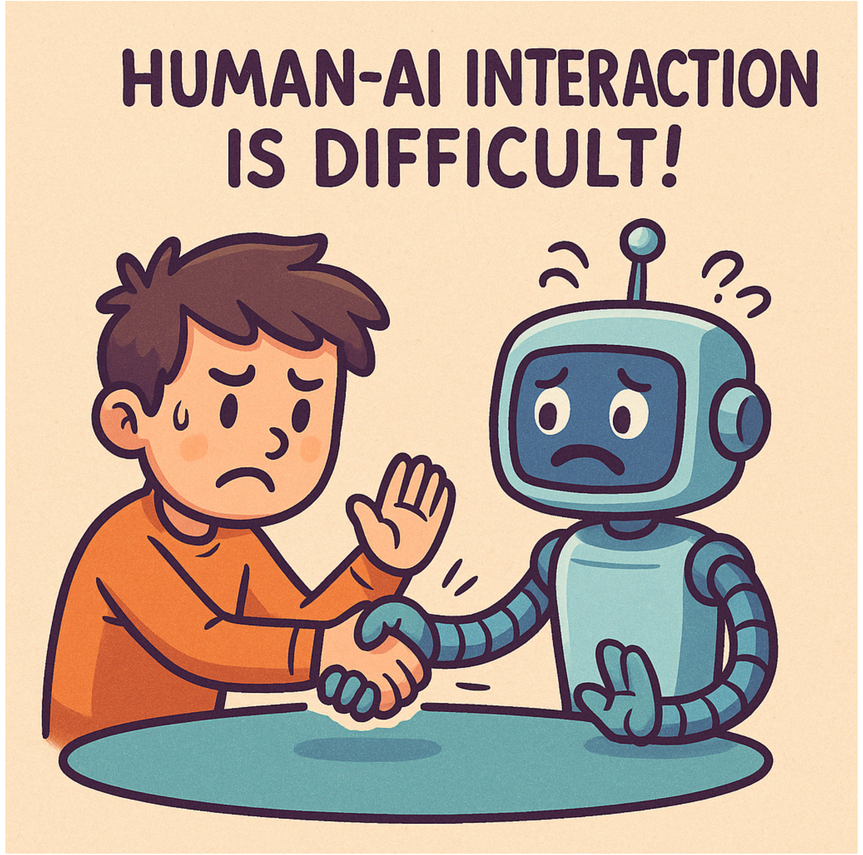
Operations defined by prompts

Operations executed by LLMs

I will call this “semantic data processing”

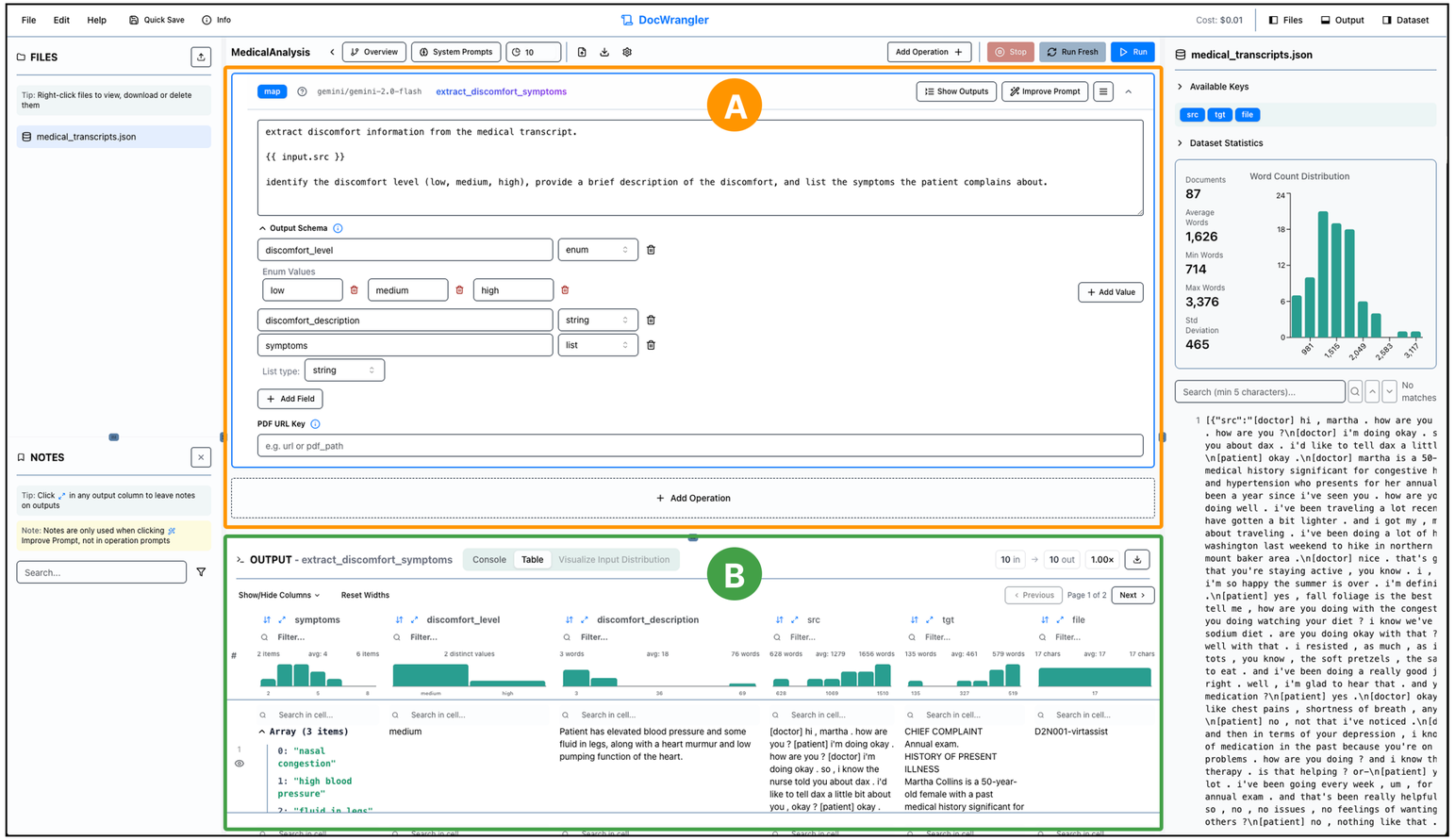
DocETL is great but not sufficient

- 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.



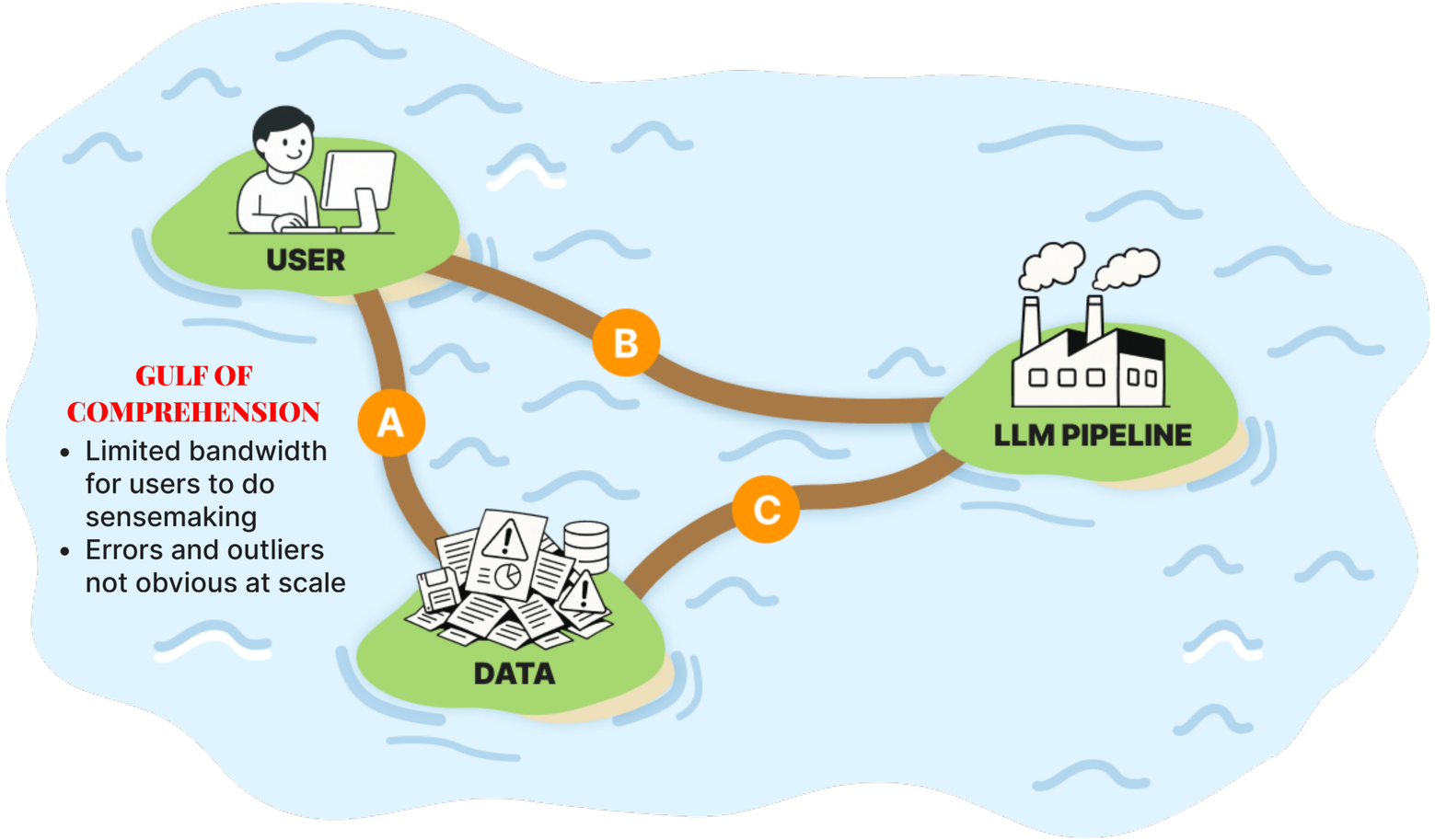
DocWrangler: An IDE for Semantic Data Processing

docetl.org/playground



- 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.”

→

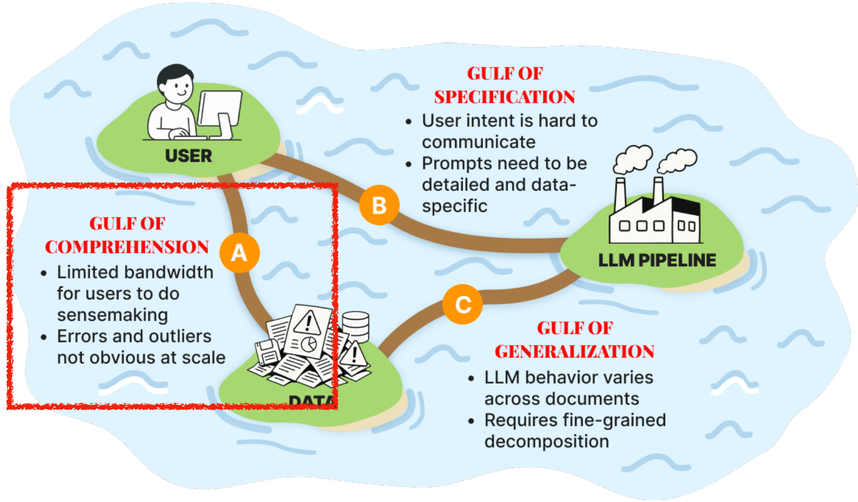
👤 **Human Refinement**
“Also explicitly indicate if any efficiency metrics (CO2 reduction) are **N/A** in the document.”

KEY ISSUE

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

Solution: In-Situ User Notes

Bridging the Comprehension Gulf



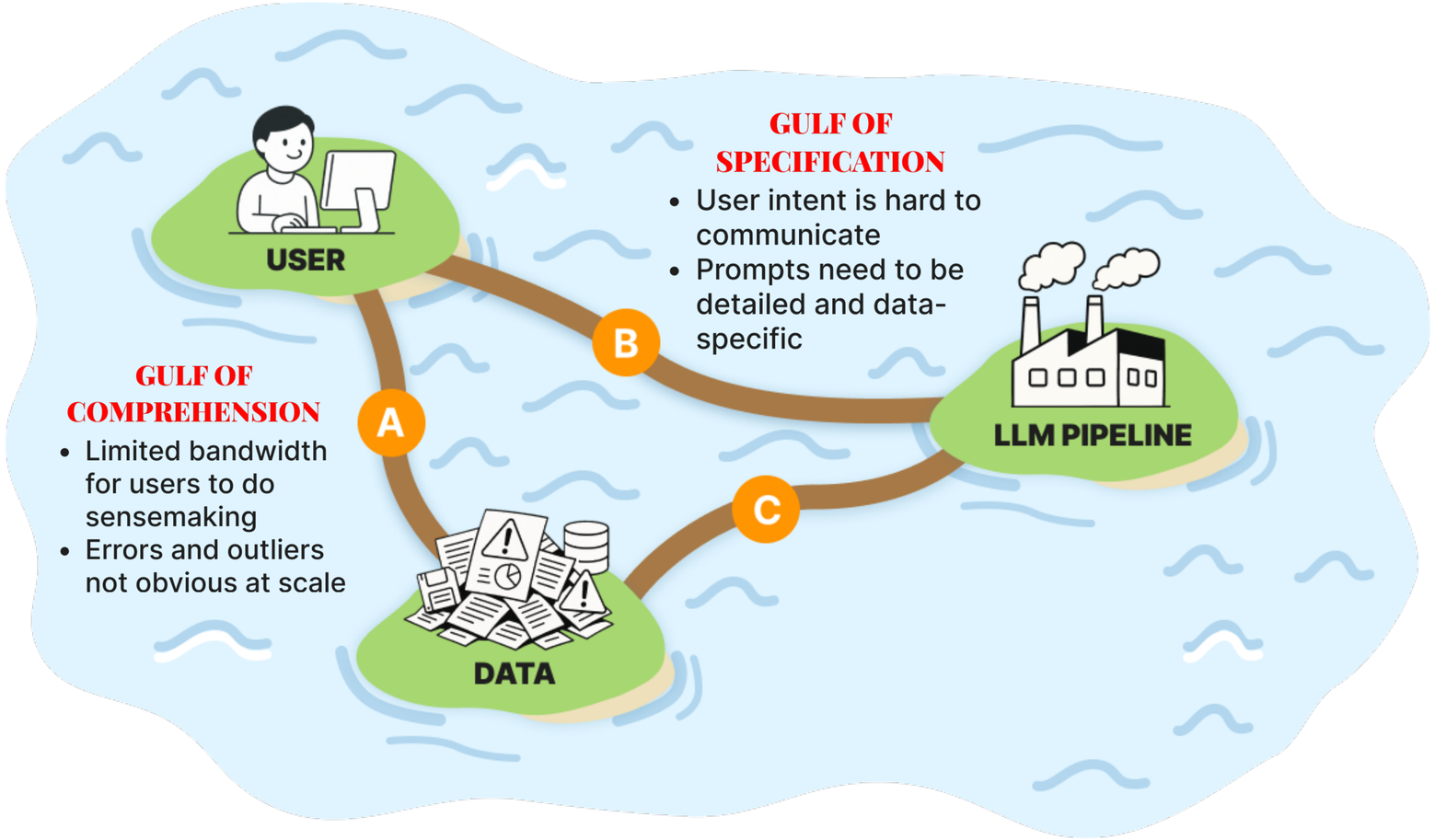
- DocETL user challenges
  - Keep track of notes on outputs, documents, and LLM behavior
  - Find patterns in data and behavior

KEY IDEA

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

Screenshots: In-Situ User Notes

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."

→

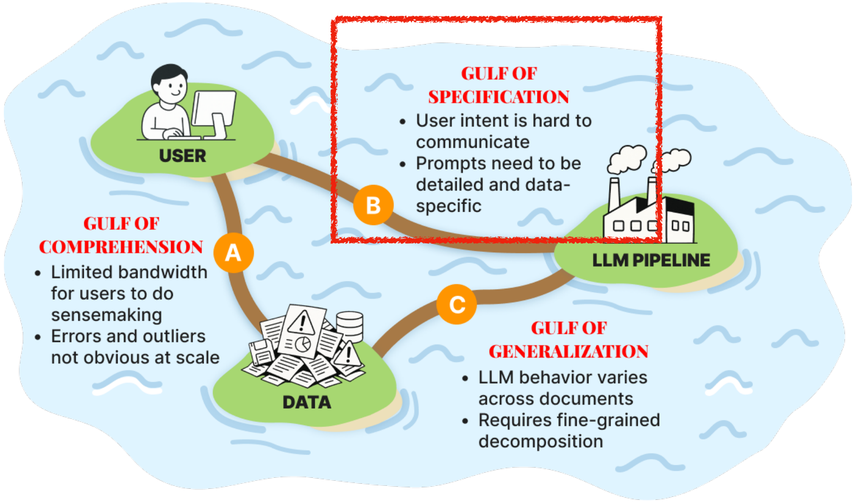
👤 **Human Refinement**
"Add the climate **impact metric** - what percentage carbon reduction was achieved through this intervention?”

KEY ISSUE

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

Solution: Prompt Refinement

Bridging the Specification Gulf

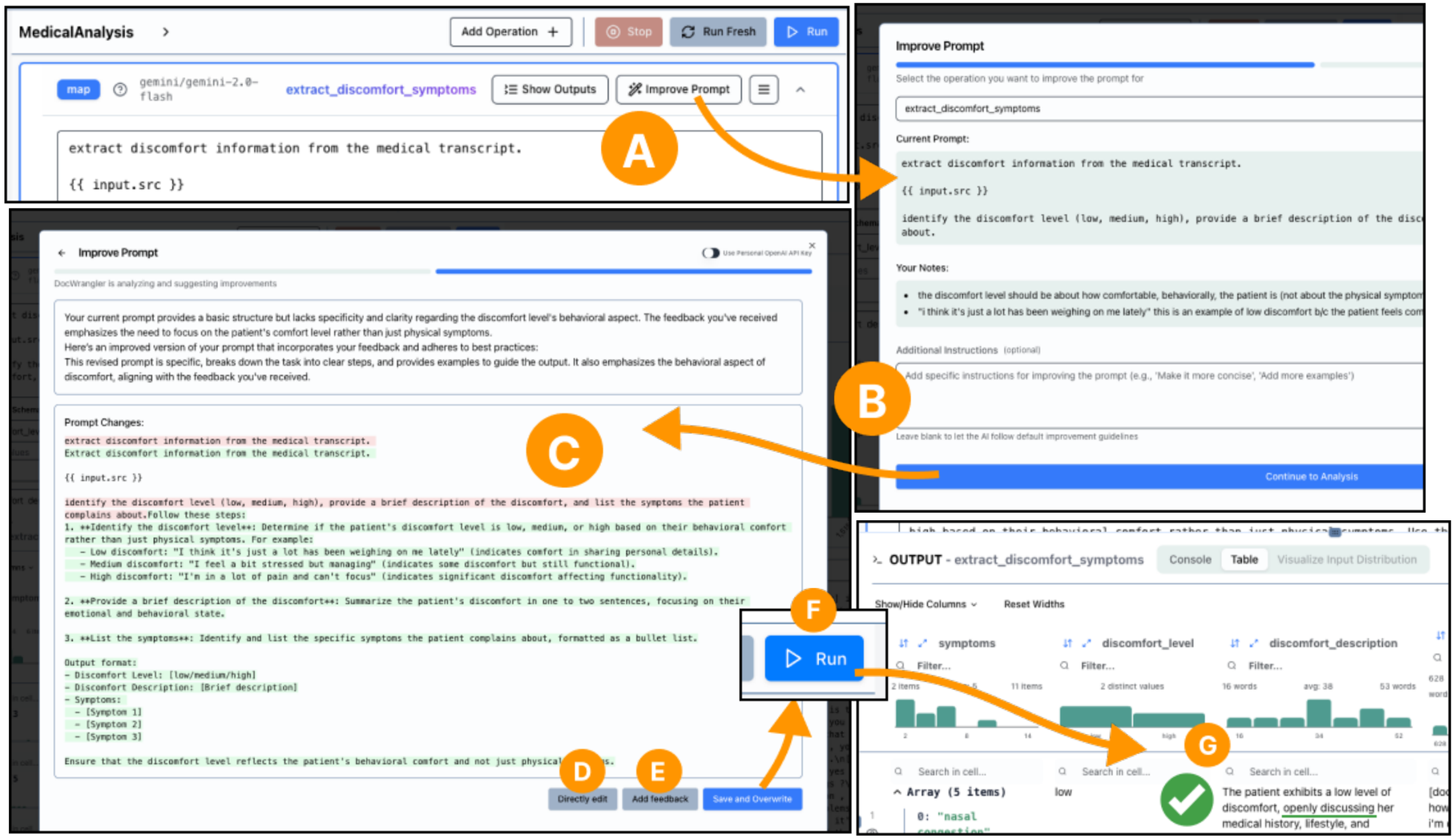


- DocETL user challenges
  - Translate observations into actual prompt modifications
  - Adding specificity; maybe even changing the task altogether

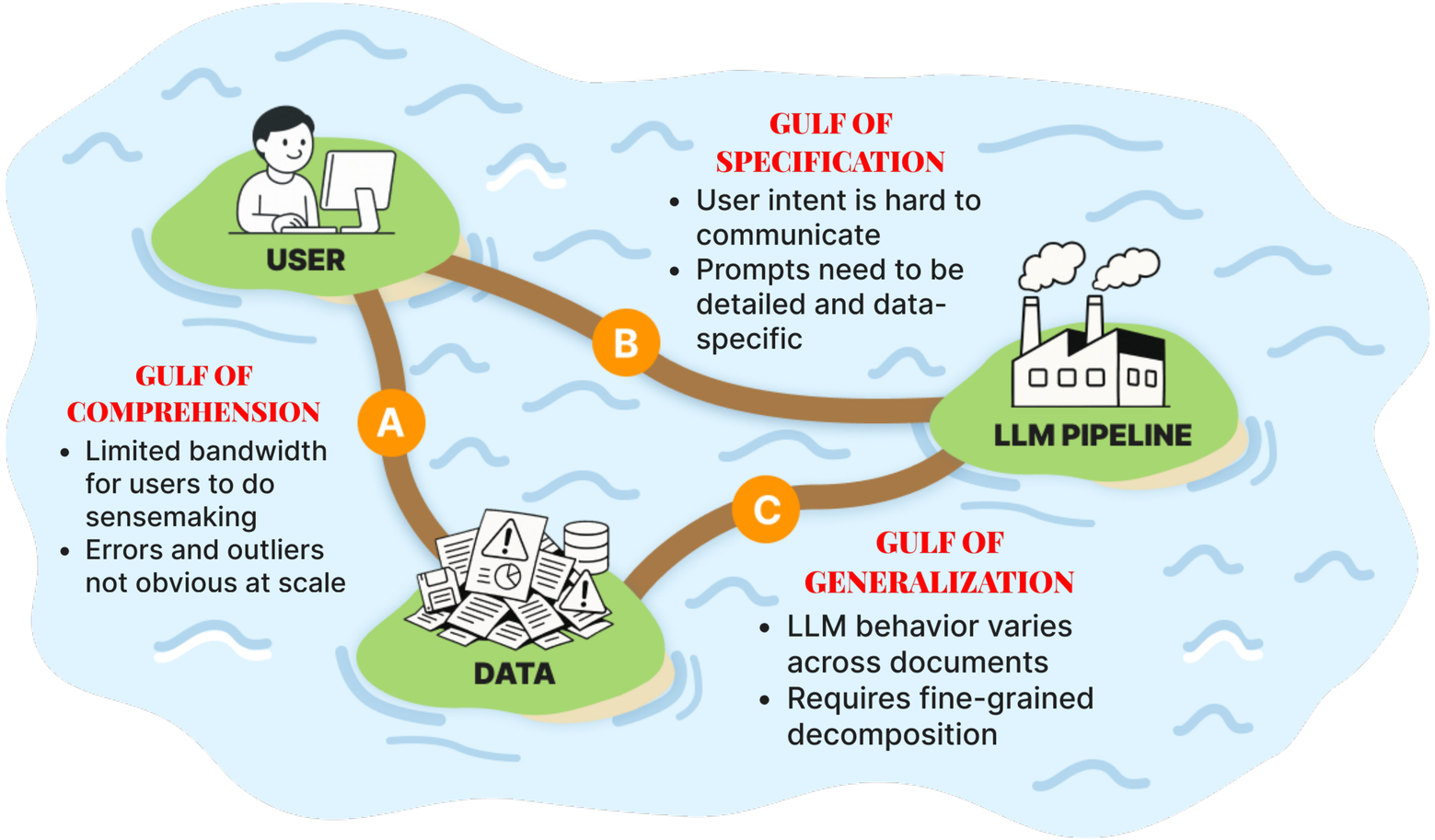
KEY IDEA

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

Screenshots: Assisted Prompt Refinement

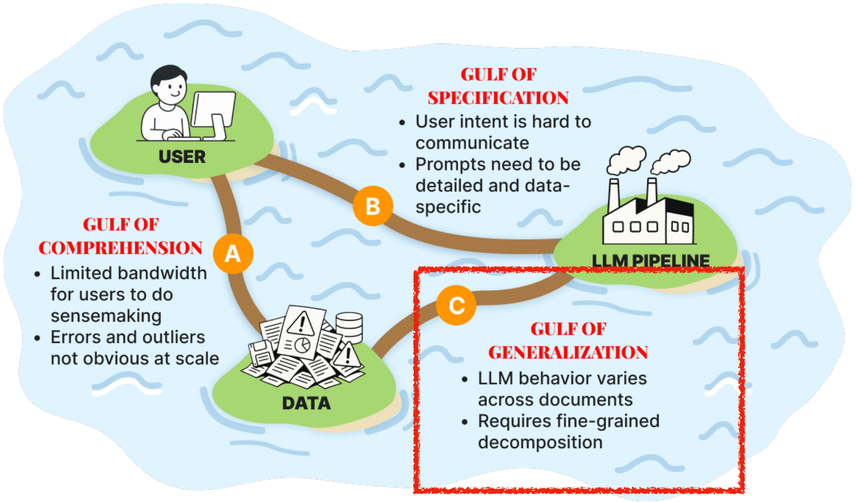


Interaction Challenges in Semantic Data Processing



Solution: Assisted Decomposition

Bridging the Generalization Gulf



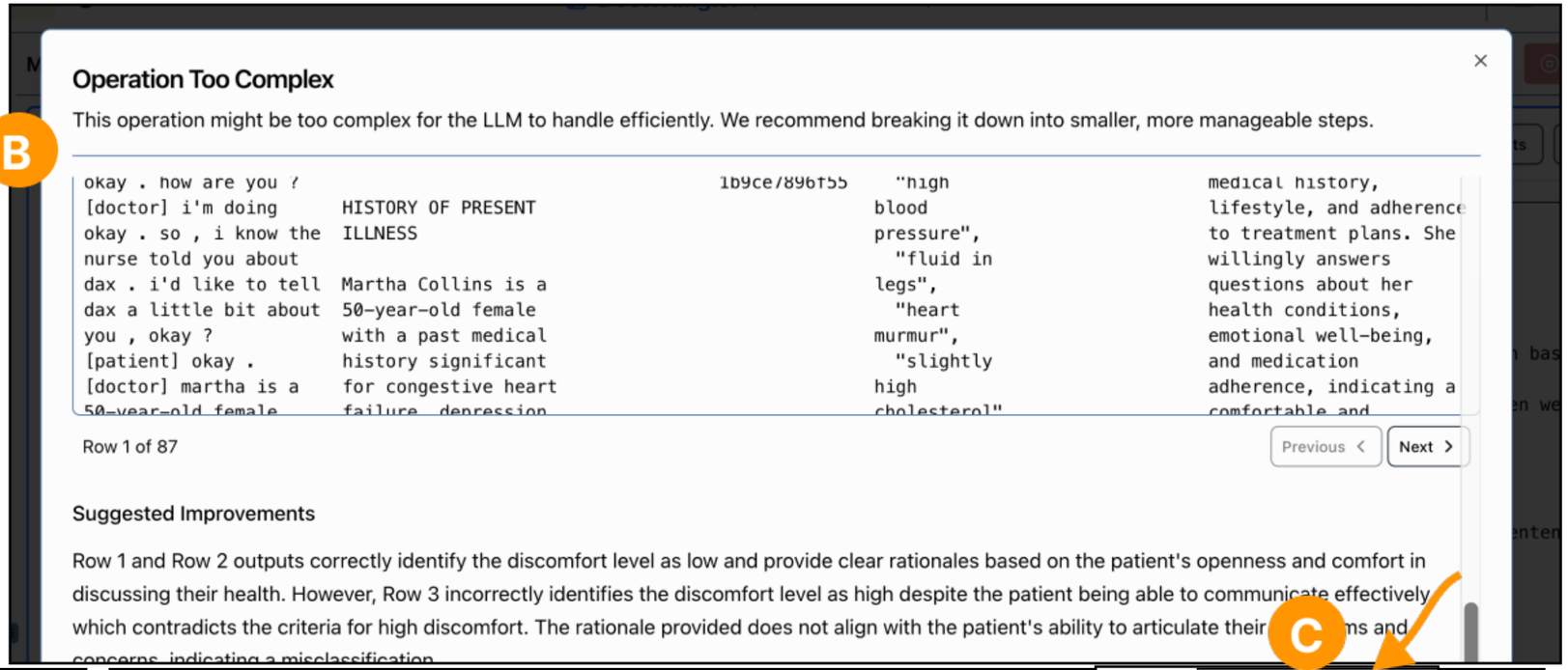
- DocETL user challenges
  - Knowing when an operation is “too complex” as-is
  - Knowing how to decompose operations

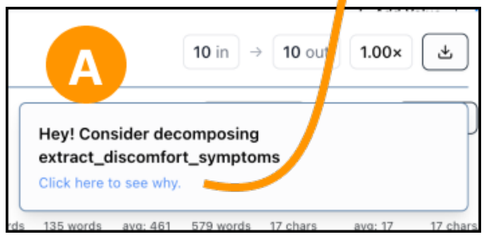
KEY IDEA

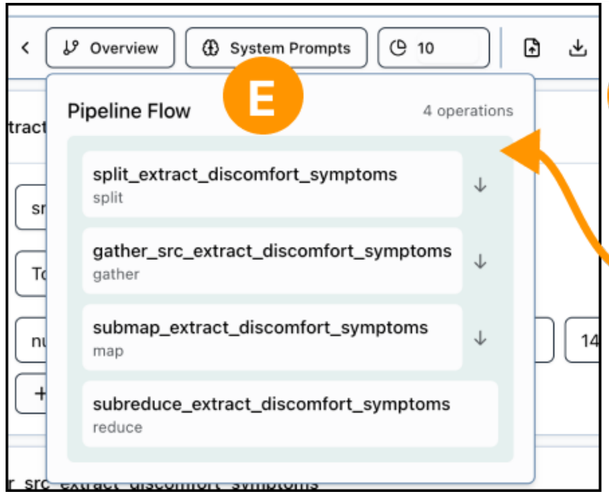
We “notify” users when we think operations are too complex & why.

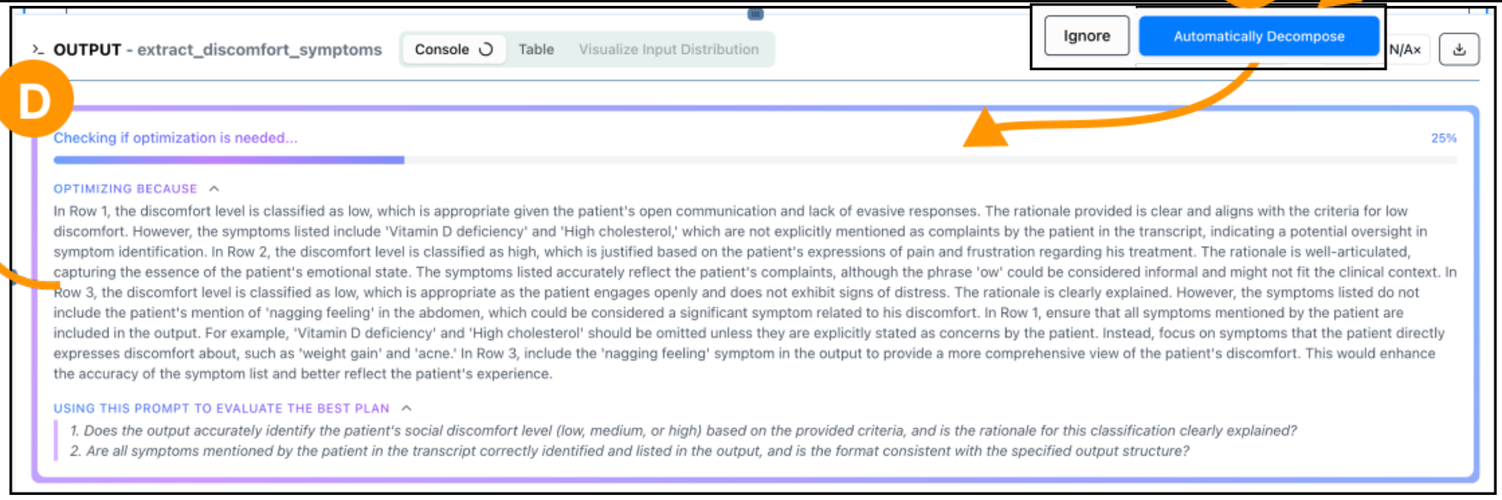
Screenshots: Operation Decomposition

*It might take 5+ minutes to run interactively

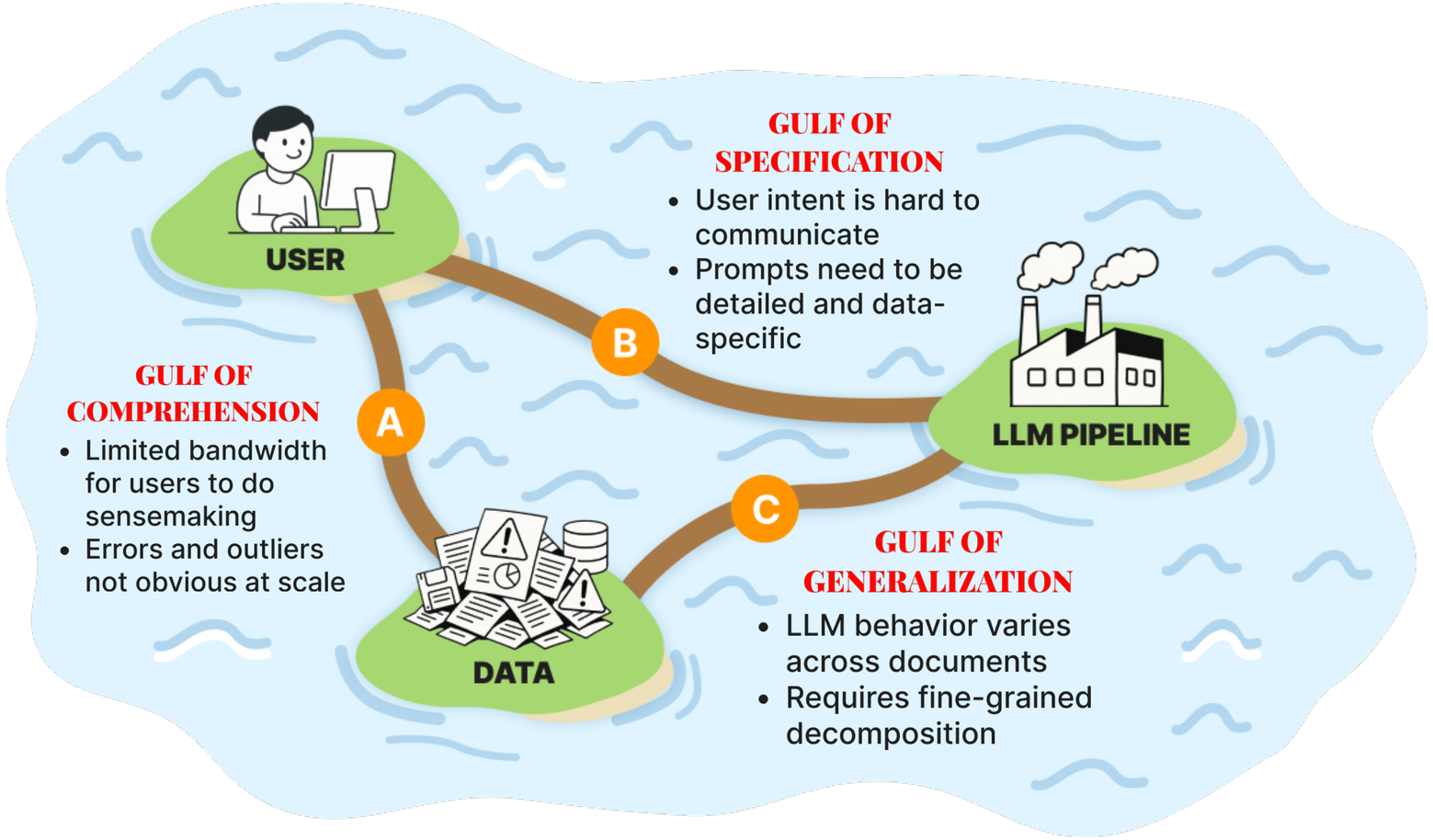


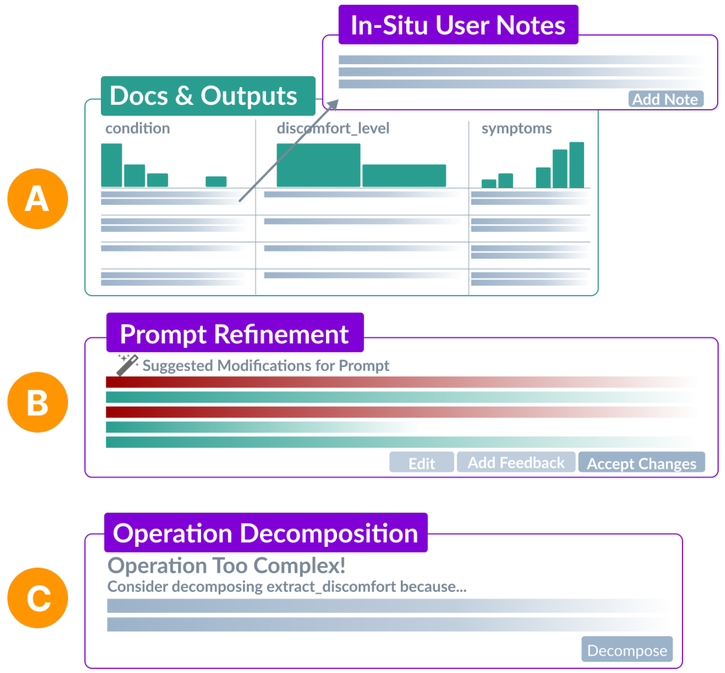






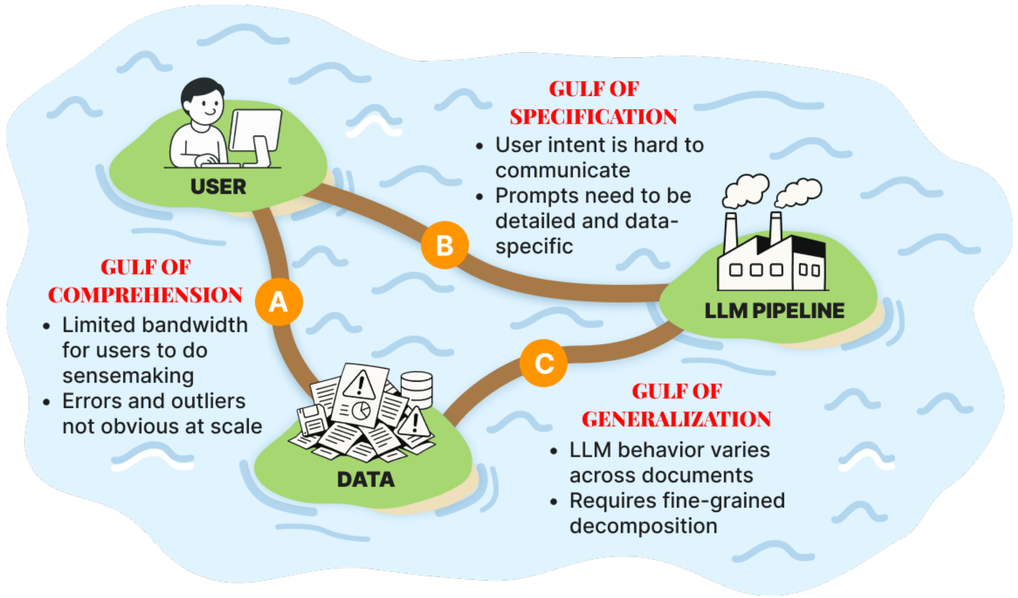
Interaction Challenges in Semantic Data Processing





Lessons Learned From User Studies

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

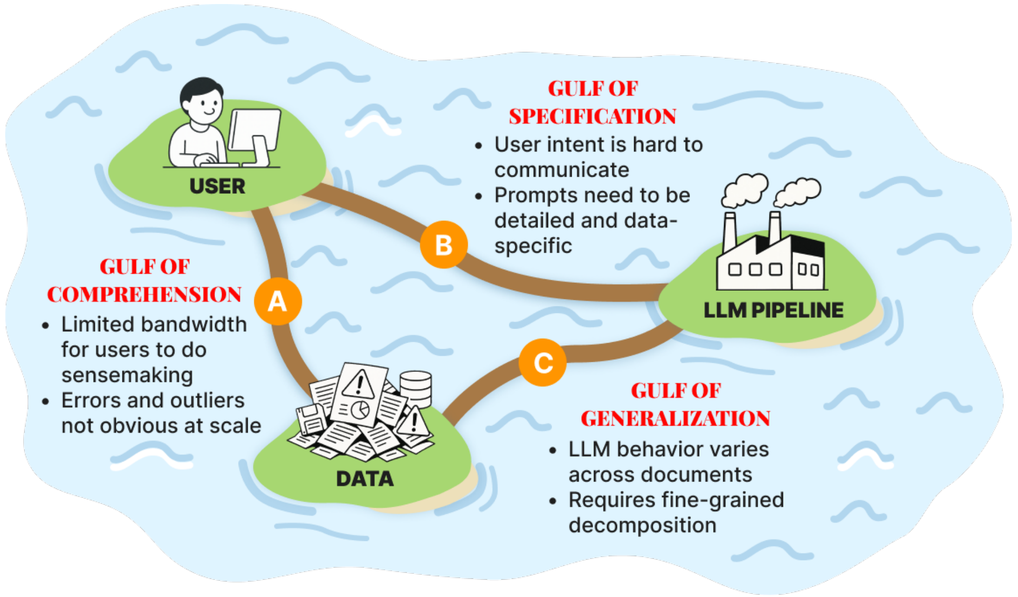


Common strategies

- 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