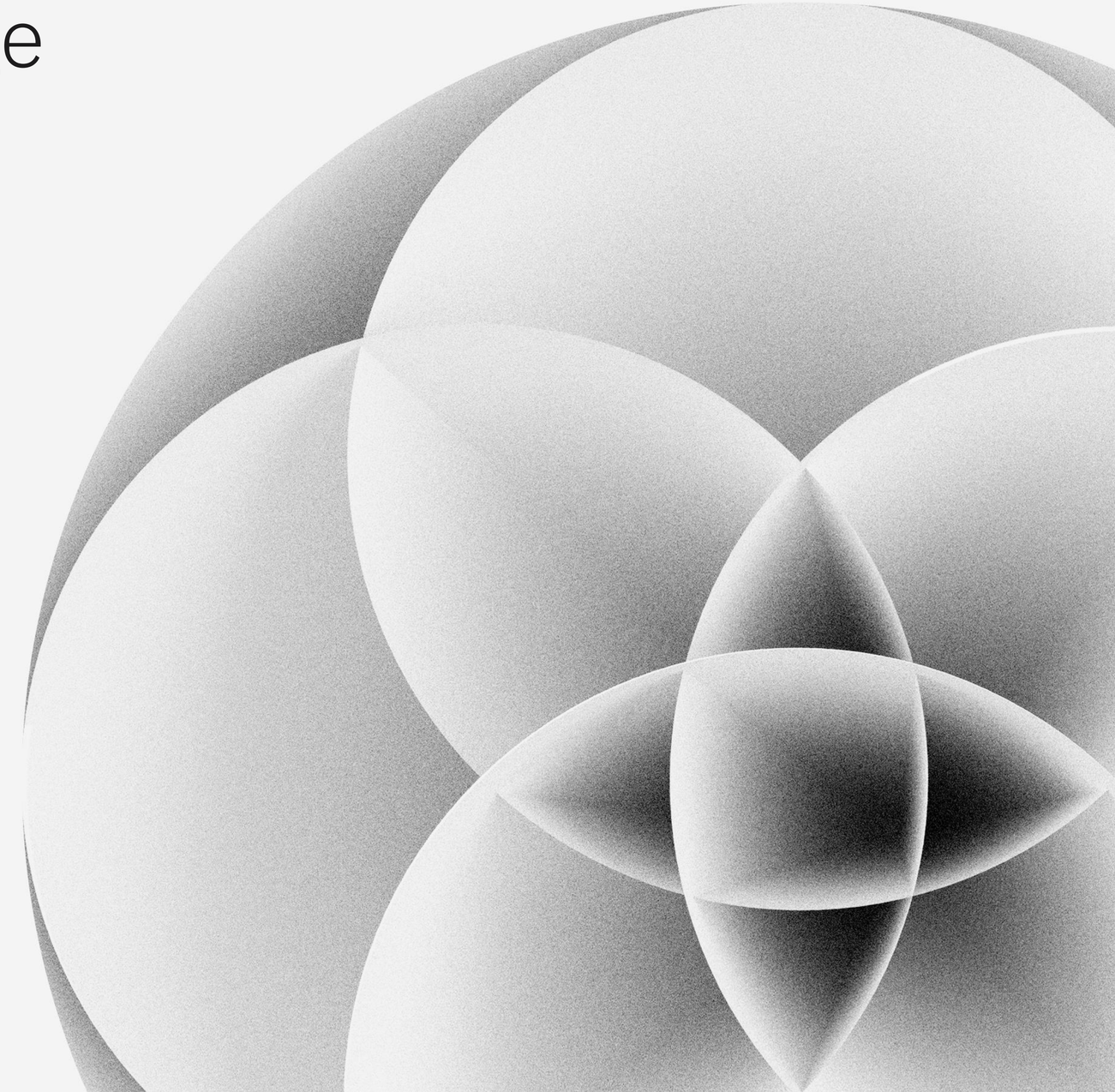
IBM Research

Prompt Declaration Language
PDL



Mandana Vaziri

IBM

PDL Team

Mandana Vaziri
Principal Research Scientist
—
mvaziri@us.ibm.com

Louis Mandel
Research Scientist
—
lmandel@us.ibm.com

Claudio Spiess
PhD Candidate, UC Davis
—
cvspiess@ucdavis.edu

Martin Hirzel
Principal Research Scientist
—
hirzel@us.ibm.com

IBM T.J. Watson Research Center
Yorktown Heights, NY

Prompt engineering is hard

How does PDL help?

Prompts at the forefront

Every character counts

PDL written in YAML

Single declarative language with control structures, and functions for pattern reuse

Few orthogonal features

Composition of LLMs and code

Need to chain LLMs and tools

PDL abstracts away the plumbing necessary for such compositions

Supports a wide variety of model providers and models, based on LiteLLM

Implicit accumulation of messages

LLMs accept as input a structured list of messages

PDL keeps track of the context implicitly, making programs much less verbose

Support for chat APIs

Type checking

Often LLM input and outputs have unchecked data formats

PDL provides type checking of both input and output of models. Types feed seamlessly into constrained decoding

Intrinsics

LLM outputs can contain hallucinations

PDL is based on granite-io and supports the following intrinsics:

Thinking
Hallucinations
Answerability
Certainty
Citations
Query-rewrite

Automated parallelization

Often LLM calls are slow

In PDL, all model calls are asynchronous and will be executed in parallel in the absence of data dependencies

Automated Prompt Optimization

Need for prompt tuning

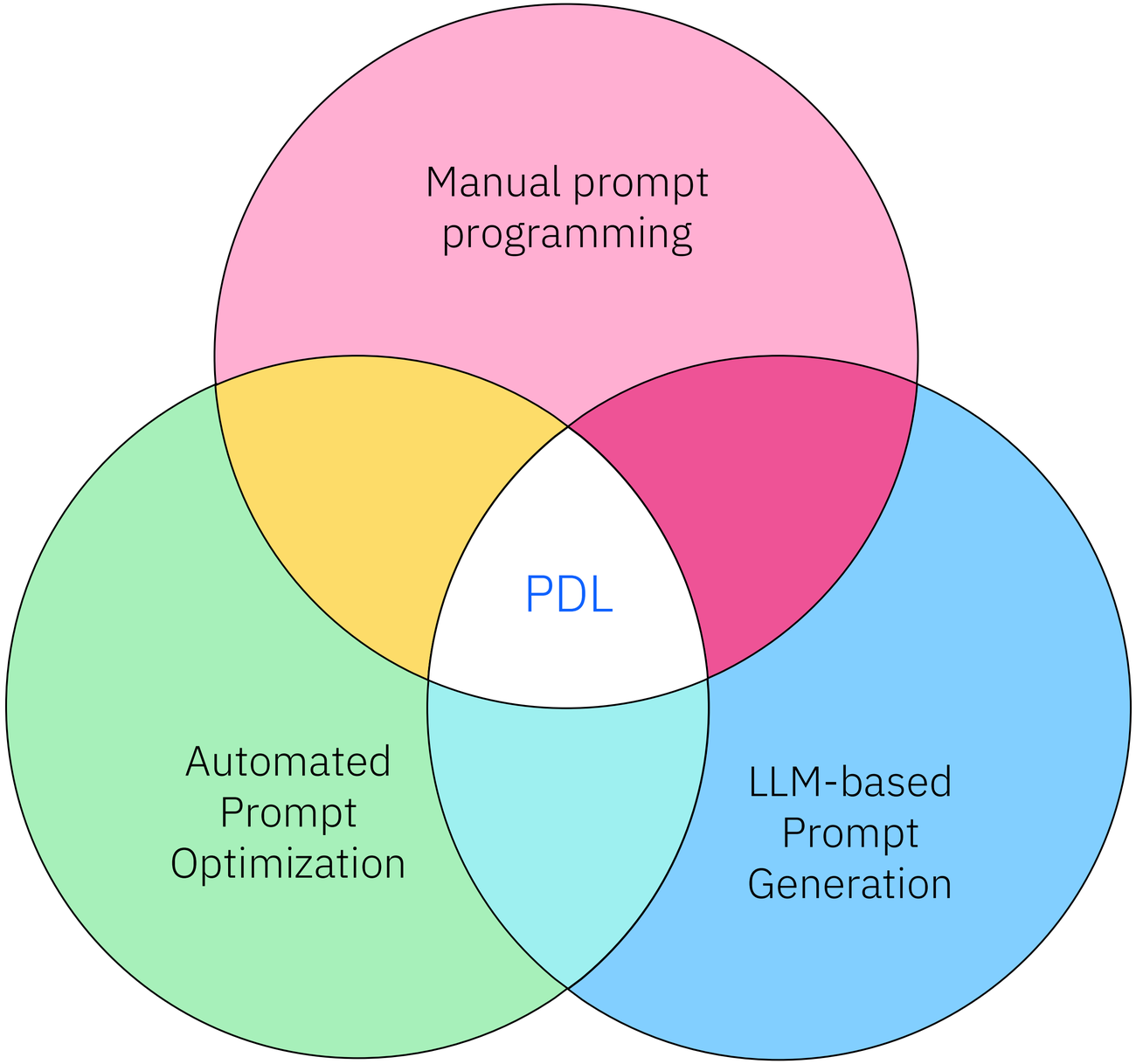
AutoPDL starts with a PDL program with variables, a domain specification, and a dataset. It automatically finds optimal values for the variables

AutoPDL can be used to optimize prompting patterns and few-shots

One Representation Many Uses

PDL can be used for low-level prompt programming and manual prompt pattern customization.

Its declarative nature makes it amenable to automated prompt optimization (AutoPDL), and to be generated effectively by LLMs.



Demonstration Links

See Demo!

Early Adopters

IBM AI Agent SWE-1.0

IBM CISO Compliance Agent

Learn More

My Contact Information:

mvaziri@us.ibm.com

PDL Git Repo

https://github.com/IBM/prompt-declaration-language

Try the PDL tutorial and the examples today!

Give us feedback!

PDL Paper

https://arxiv.org/abs/2410.19135

Read about PDL!

Check out PDL at ICML/PRAL!
https://pral-workshop.github.io

AutoPDL Paper

https://arxiv.org/abs/2504.04365

Read about AutoPDL!

PDL Chatbot Example

```yaml
text:
- read:
  contribute: [context]
  message: |
    What is your query?
- repeat:
    text:
      - model: watsonx/ibm/granite-13b-chat-v2
        parameters:
          stop: ["\n\n"]
      - def: question
        read:
        contribute: [context]
        message: |

          Enter a query or say "quit" to exit.
  until: ${question == "quit"}
```

(a) Code

% pdl chatbot.pdl

What is your query?

What's a language salad?

A language salad is a term used to describe a mix of different languages and dialects in a single conversation or piece of text. It can be seen as [...]

Enter a query or say "quit" to exit.

Say it as a poem!

In a world where many tongues are sown,
A language salad is born, in joy they're grown.
A medley of words, in harmony flow,
Swirling colors of speech, in a vibrant show.

Enter a query or say "quit" to exit.

quit

(b) Interpreter trace

Support for IBM Granite Intrinsics

Intrinsics are special meta-data that help qualify the output of a model.
TODO: other models
Supported intrinsics:

Thinking
Hallucinations
Answerability
Certainty
Citations
Query-rewrite

```yaml
description: GraniteIO hallucination example
defs:
  doc:
    data:
      text: |
        Audrey Faith McGraw (born September 21, 1967) is an American singer
        ...

text:
- Did Faith Hill take a break from recording after releasing her second album, It Matters to Me?
- processor:
    model: "granite3.2:2b"
    backend: openai
  parameters:
    documents:
    - ${ doc }
    controls:
      hallucinations: true
  modelResponse: output
- "\nHallucinations:\n"
- for:
    hallucination: ${ output.results[0].next_message.hallucinations }
  repeat:
    text:
    - "Hallucination Risk: ${ hallucination.risk }"
    - "\nSentence: ${ hallucination.response_text }"
  join:
    with: "\n"
```

AutoPDL Results

Start from a dataset and a combinatorial space of agentic and non-agentic prompting patterns. AutoPDL automatically picks few-shot samples, instructions, and a pattern.

Paper at AutoML'25

https://arxiv.org/pdf/2504.04365

Table 1: Model accuracies across datasets for baseline (zero-shot) and optimized versions.

Dataset	Model	Accuracy			Pattern	Runtime
		Zero-Shot	Optimized	Delta		
FEVER	Granite 3.1 8B	78.3 %	79.0 %	+0.7pp	ReWOO (5 shot)	08:55
	Granite 13B Instruct V2	6.5 %	75.4 %	+68.9pp	ReWOO (3 shot)	08:12
	Granite 20B Code	39.7 %	64.2 %	+24.5pp	CoT (3 shot)	05:06
	Granite 34B Code	56.4 %	65.6 %	+9.2pp	CoT (3 shot)	03:47
	LLaMA 3.1 8B	68.5 %	78.0 %	+9.5pp	CoT (3 shot)	05:24
	LLaMA 3.2 3B	38.0 %	66.9 %	+28.9pp	ReWOO (5 shot)	09:08
	LLaMA 3.3 70B	67.6 %	77.5 %	+9.9pp	ReWOO (5 shot)	09:32
GSM8K	Granite 3.1 8B	74.2 %	(74.2 ± 0.6) %	+0.0pp	Zero-Shot (Baseline)	08:56
	Granite 13B Instruct V2	23.0 %	(30.9 ± 1.0) %	+7.9pp	CoT (3 shot)	09:20
	Granite 20B Code	68.7 %	(68.7 ± 0.1) %	+0.0pp	Zero-Shot (Baseline)	09:27
	Granite 34B Code	72.1 %	(72.1 ± 0.1) %	+0.0pp	Zero-Shot (Baseline)	08:52
	LLaMA 3.1 8B	78.4 %	(85.3 ± 0.6) %	+6.9pp	CoT (5 shot)	08:48
	LLaMA 3.2 3B	71.8 %	(75.3 ± 0.4) %	+3.5pp	CoT (3 shot)	16:36
	LLaMA 3.3 70B	85.5 %	(95.4 ± 0.2) %	+9.9pp	CoT (3 shot)	07:50
MBPP+	Granite 3.1 8B	62.9 %	(62.9 ± 0.0) %	+0.0pp	Zero-Shot (Baseline)	02:14
	Granite 13B Instruct V2	10.7 %	(19.2 ± 1.2) %	+8.5pp	ReAct (5 shot)	04:02
	Granite 20B Code	51.8 %	(51.8 ± 0.4) %	+0.0pp	Zero-Shot (Baseline)	03:43
	Granite 34B Code	48.7 %	(61.3 ± 1.0) %	+12.6pp	ReAct (3 shot)	04:54
	LLaMA 3.1 8B	61.2 %	(62.8 ± 4.0) %	+1.6pp	ReAct (5 shot)	01:45
	LLaMA 3.2 3B	58.0 %	(58.0 ± 0.4) %	+0.0pp	Zero-Shot (Baseline)	02:01
	LLaMA 3.3 70B	71.4 %	(71.4 ± 0.0) %	+0.0pp	Zero-Shot (Baseline)	02:27