AESHM Data Analytics Lab Seminar

Customizing LLMs for Academic Research

Jewoo Kim, PhD & Chunsheng (Jerry) Jin, PhD Candidate

IOWA STATE UNIVERSITY

IOWA STATE UNIVERSITY

AESHM Data Analytics Lab

- Launched in 2020.
- Activity
  - Data collection for AESHM graduate research
  - Data Analytics Lab Seminar
    - 1st Seminar: GenAI Applications to Research (Fall 2024)
    - 2nd Seminar: GenAI for Hospitality Management & Applied Statistics

GenAI Inherent Risks

- Hallucination
  - Incorrect responses in a confident tone.
  - Happens when models' pre-trained knowledge does not support accurate response generation.
    - Biased information
    - Outdated information
    - Limited information

GenAI Inherent Risks

- Bias
  - Over-representation
    - Reproduces stereotypes and unfounded generalizations.
  - Driving factors
    - Unbalanced pre-training datasets
    - LLM applications to domains where relevant data is lacking

Iowa State University

LLM Applications in Research

Validation of domain knowledge

- Confirms whether LLMs have sufficient domain knowledge to reliably conduct research tasks.

Model enhancement in domain-specific tasks

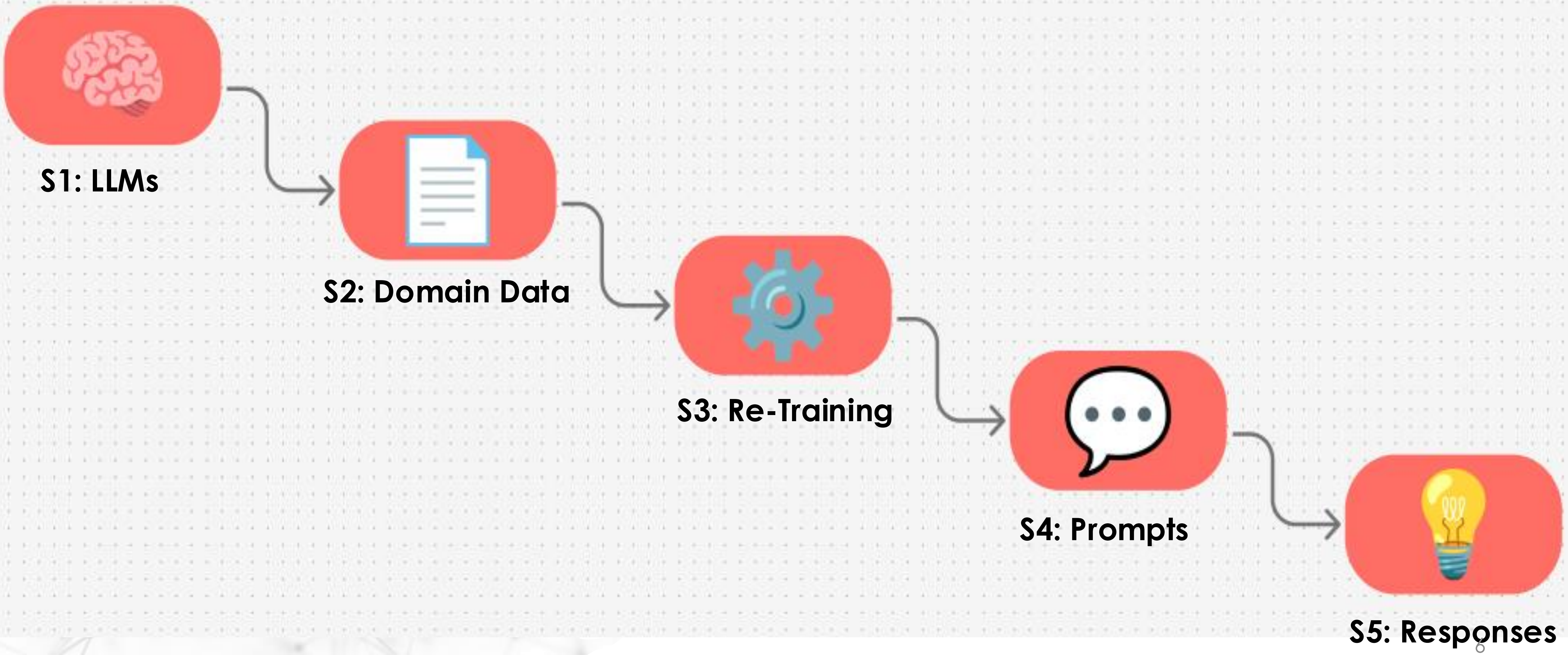
- Prompt engineering
  - Designs and refines prompts to generate desired outputs.
  - Ex) Self-reflection (SR), Chain-of-Thought (CoT)

LLM Applications in Research

Model enhancement in domain-specific tasks

- Model adaptation
  - Customizes LLMs to enhance performance in domain- or task-specific contexts.
- Adaptation strategies
  - Fine-Tuning
  - In-Context Learning (ICL)
  - Retrieval-Augmented ICL (RA-ICL)

Fine-Tuning


S1: LLMs
S2: Domain Data
S3: Re-Training
S4: Prompts
S5: Responses

Fine-Tuning

Model customization

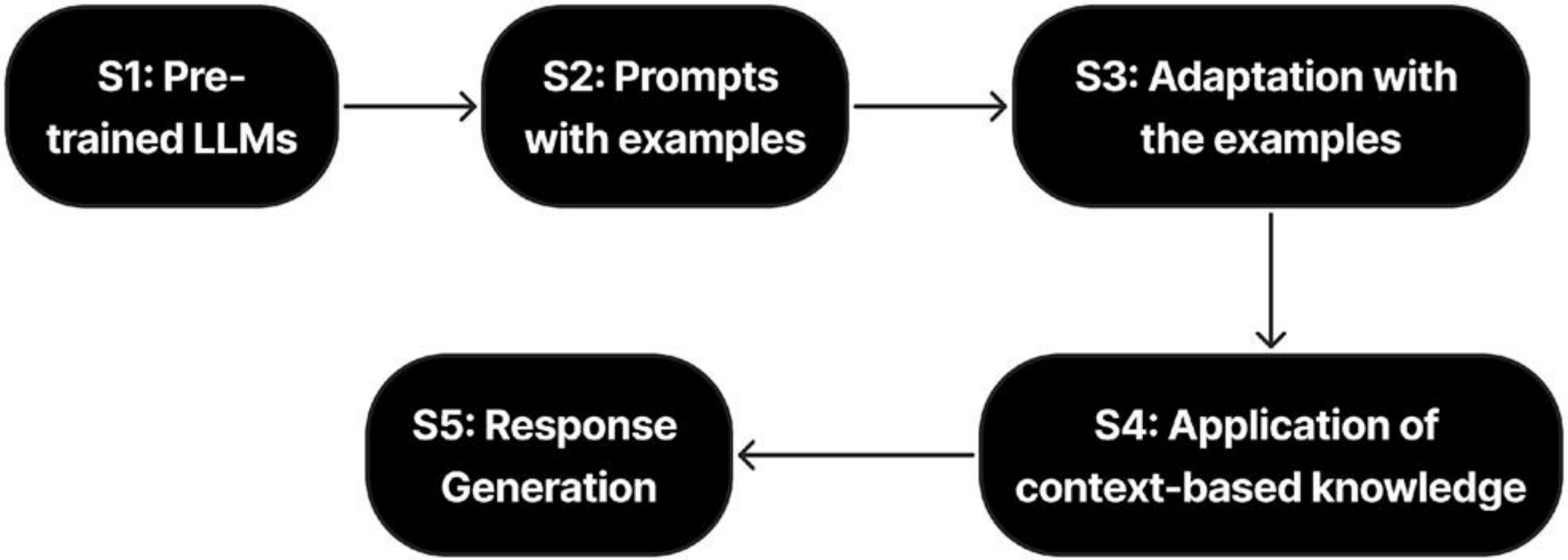
Tailors pre-trained LLMs by further training them with task- or domain-specific datasets.

Parametric adaptation

Updates model parameters to better reflect domain knowledge and reduce reliance on pre-training data.

In-Context Learning (ICL)

- Few-shot learning



In-Context Learning (ICL)

Non-parametric adaptation

- Provides contexts within prompts to guide models.
- Prompts can include only task instructions (zero-shot) or both task instructions and examples (few-shot).

Meta-learning (Learning to learn)

- During pre-training, LLMs learns an internal mechanism to infer new tasks from prompts.

Retrieval-Augmented ICL (RA-ICL)

Challenge in ICL	How to select the most relevant examples for ICL? • Manual selection in basic ICL • Systematic selection in RA-ICL
RAG + ICL	Adds Retrieval-Augmented Generation (RAG) to bring in relevant examples from external databases for ICL.

Retrieval-Augmented ICL (RA-ICL)

- How RAG works

Relevant information retrieval

Adaptation through context augmentation

Grounded response generation

Retrieval-Augmented ICL (RA-ICL)

- Advantages of grounded response generation

Higher Accuracy

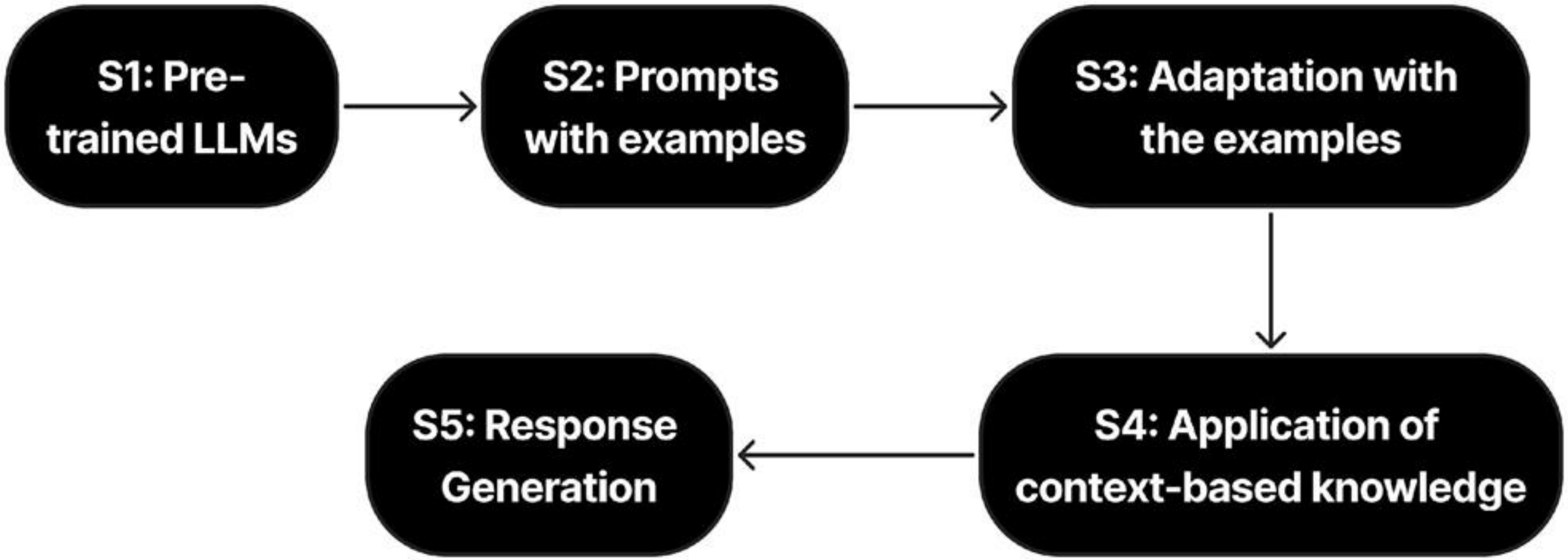
Greater transparency

Reduced hallucination & bias

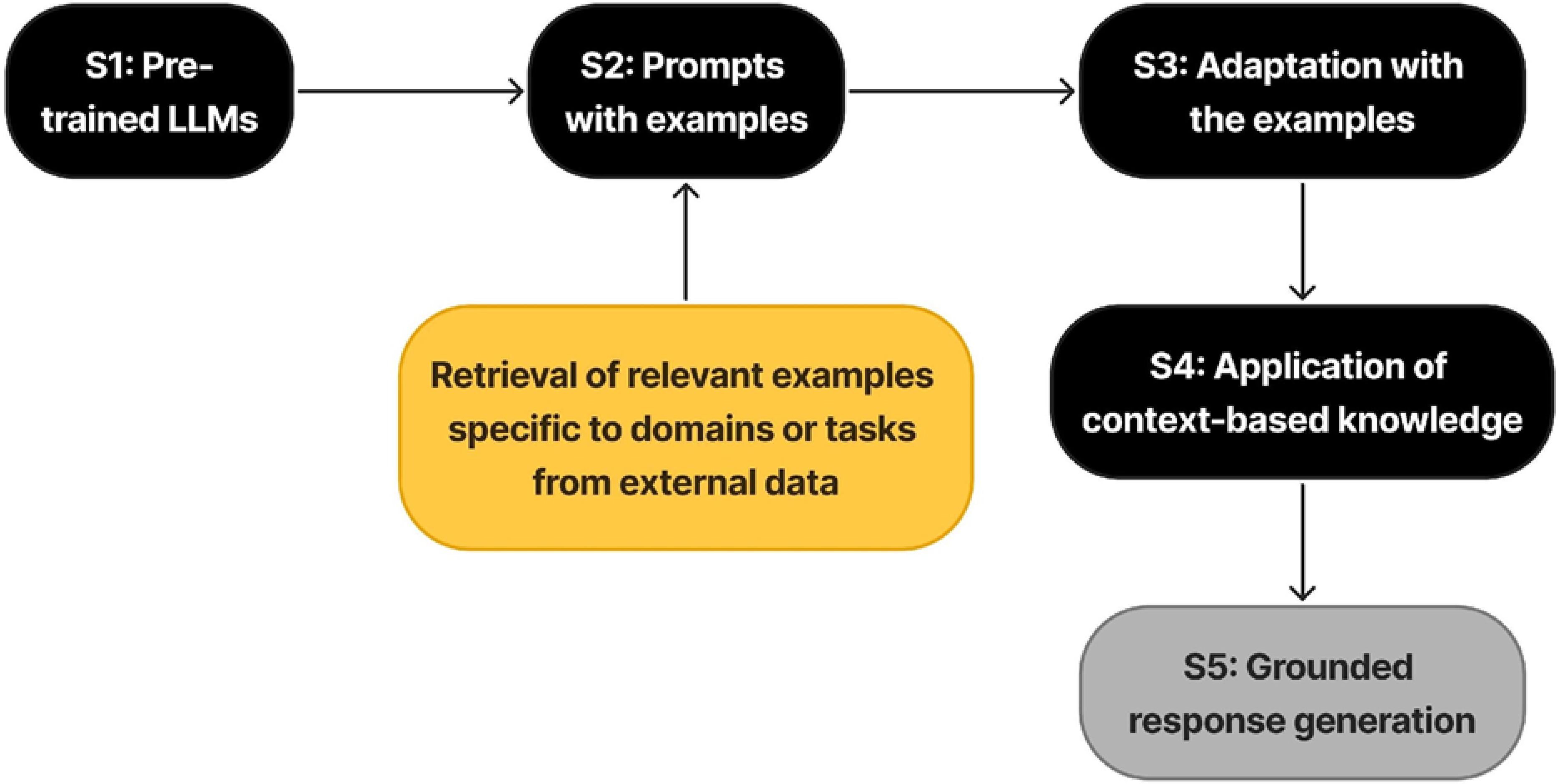
Increased trust

In-Context Learning (ICL)

- Few-shot learning



RA-ICL Process



Empirical Test

- Dataset: online hotel reviews
  - Test set - 548 reviews
  - Training set – 2,414 reviews
- Sentiment analysis
  - Evaluates sentiments of four lodging aspects
    - Value, room, location, & service

Empirical Test: ICL

- GPT API was used to evaluate sentiment across four aspects for each review.
- Without any additional model training, API was directly applied to the test dataset.
- Step-by-step instructions and review examples with sentiment scores were included in the prompt to guide model output.

Empirical Test: ICL

- Evaluation steps

```
prompt = f"""
Let's think step by step.
step1: The 'Grading Rubrics' below show example reviews with the ratings (1 to 5) of four aspects (location, room, service, value);
step2: The goal is to analyze and assess the provided hotel review content. Review the hotel reviews and determine if they mention a
    - Mentions of price/what we got for what we paid ⇒ value (not service).
    - Staff behavior, check-in/out efficiency, responsiveness ⇒ service.
    - Cleanliness, bed, noise in room, amenities inside the room ⇒ room.
    - Proximity, transportation, neighborhood, safety ⇒ location.
step3: Assign a rating between 1 and 5 for each aspect. If a specific aspect is not addressed in the review, infer the aspect rating
step4: If you have only performed the rating assignment for some of the apects, implement the previous steps until a rating is assig
```

Empirical Test: ICL

- Example reviews and their ratings for each aspect.

```
Rating Rubrics (1–5)

Location
Review: “I would not recommend this place to anyone!” => Rating: 1 (Very negative, recommending against staying)
Review: “We were moved over a mile away.” => Rating: 2 (Strong dissatisfaction / inconvenience)
Review: “Not central, but metro/taxi handled it.” => Rating: 3 (Mixed/neutral; workable alternatives)
Review: “Close to several subway stations.” => Rating: 4 (Mostly positive, convenient)
Review: “Many places within walking distance.” => Rating: 5 (Highly positive, excellent accessibility)
```

Iowa State University

Empirical Test: ICL

	Value	Room	Location	Service
Accuracy	0.52	0.56	0.61	0.70

Empirical Test: Fine-Tuning

https://drive.google.com/file/d/1_0Tk3Ter0QHnJLAzKJYkKECs9o-QRXwh/view?usp=sharing

Empirical Test: RA-ICL

https://drive.google.com/file/d/1VbhBwYubz9FtQALst9lriTFzx1pHj24j/view?usp=sharing

Applications of Customized LLMs

- Text analytics
- Image/Video analytics
- AI-generated consumer personas

Iowa State University

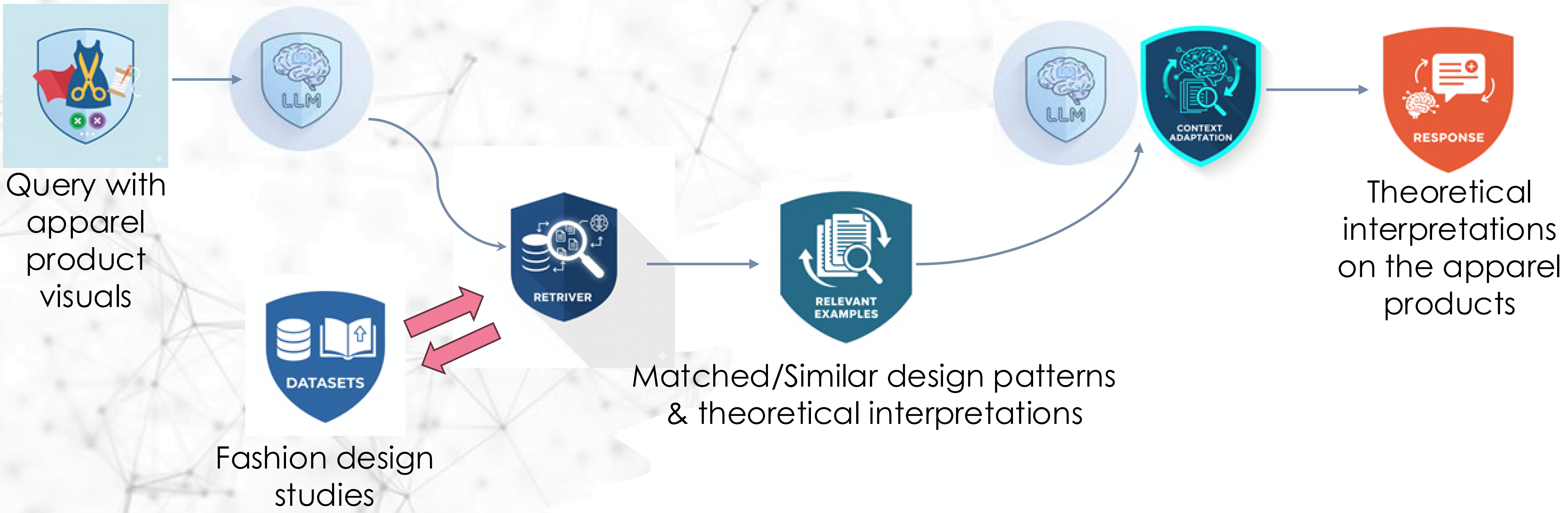
Text Analytics

- Enhanced thematic coding
  - Acts as a reliable automated coder that can apply specific, specialized coding scheme and theoretical lens.
- Reliable theme extraction
- Cross-case synthesis
  - Synthesizes thematically similar findings across different cases.

Image/Video Analytics

- Theory-based interpretations with RAG
  - Retrieves relevant studies on comparable visuals and their underlying theories from prior scholarly work.
  - Example prompt
    - "Based on the retrieved literature, interpret the symbolic meaning of asymmetry in contemporary avant-garde fashion."
  - RAG ensures the interpretations align with established scholarship.

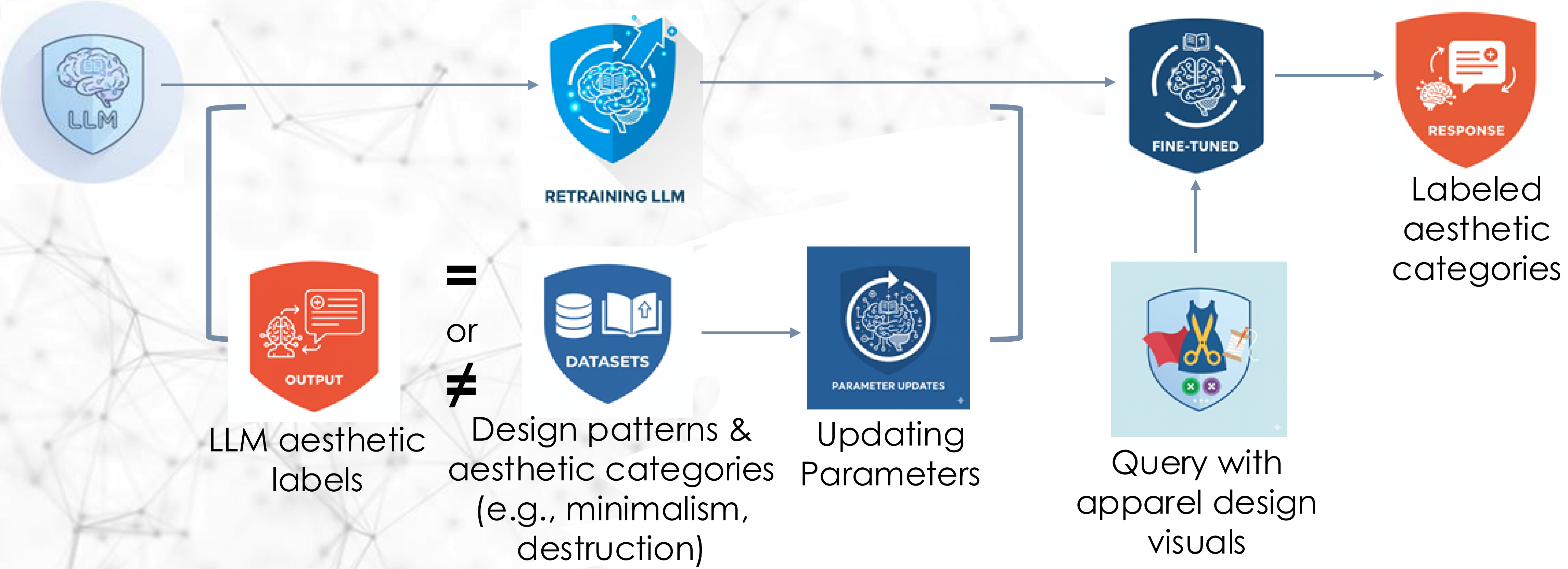
Image/Video Analytics with RAG



Image/Video Analytics

- Visual classification with Fine-Tuning
  - Retrains LLMs on coded visual categories
    - Ex) In fashion, minimalism, futuristic, nostalgic, gender-fluid
    - Learns interpretive mapping between visual cues and symbolic meanings.
  - Applies the fine-tuned classification models for qualitative coding, cross-brand comparison, and trend analysis.

Image/Video Analytics with Fine-Tuning



LLM Applications: Consumer Personas

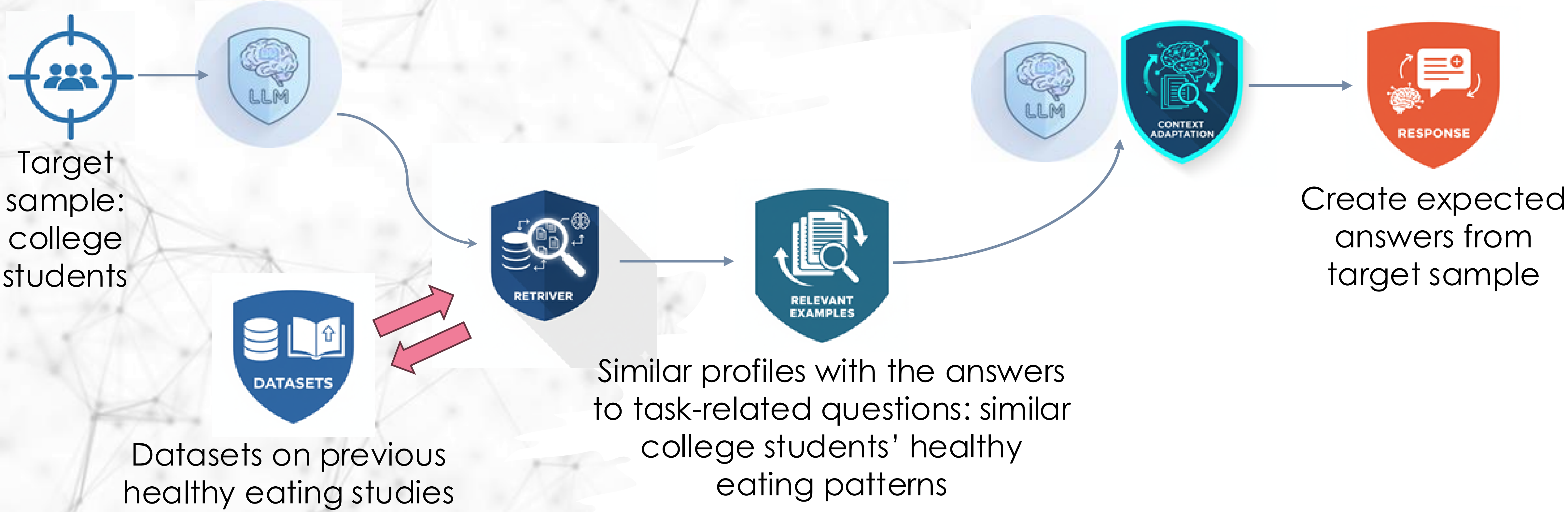
- Consumer (marketing) personas

LLM Applications: Consumer Personas

- AI-generated consumer personas
  - Use of large-scale analysis of UGC and behavioral data
    - Data sources: online reviews, social media posts, survey responses, purchase history, and browsing activity
  - Detailed, realistic profiles based on real customer data
  - Dynamic & evolving personas
    - Continue to update personas as new data become available.

LLM Applications: Consumer Personas

- Healthy eating behavior study with RAG



In Closing

- AESHM Data Analytics Courses
  - AESHM 5790 Data Analytics for AESHM
    - Statistics-based data analytics + Text analytics
    - Image & video analytics
    - Building automated GenAI agents
    - Python & LLMs
  - AESHM 4790 Business Analytics for AESHM

In Closing

- GenAI is not perfect yet.
  - Be patient to obtain the desired outcomes.
- LLMs have advanced to the current level only in three years (since late 2022).
  - What seems challenging today may become routine tomorrow.
- Do not wait! Start exploring now.

IOWA STATE UNIVERSITY

Materials for This Seminar:

https://github.com/chunshengj/AESHM-Data-Analytics-Lab-Workshop-3

Resources from Last Year's Seminar:

https://github.com/chunshengj/-Users-sheng-Jupyter-AI_Workshop_AESHM

References:

DeepLearning.AI. (2025). *Retrieval Augmented Generation (RAG)* https://www.coursera.org/learn/retrieval-augmented-generation-rag?

DeepLearning.AI. (2024). *Building Systems with the ChatGPT API*. https://www.deeplearning.ai/short-courses/building-systems-with-chatgpt/