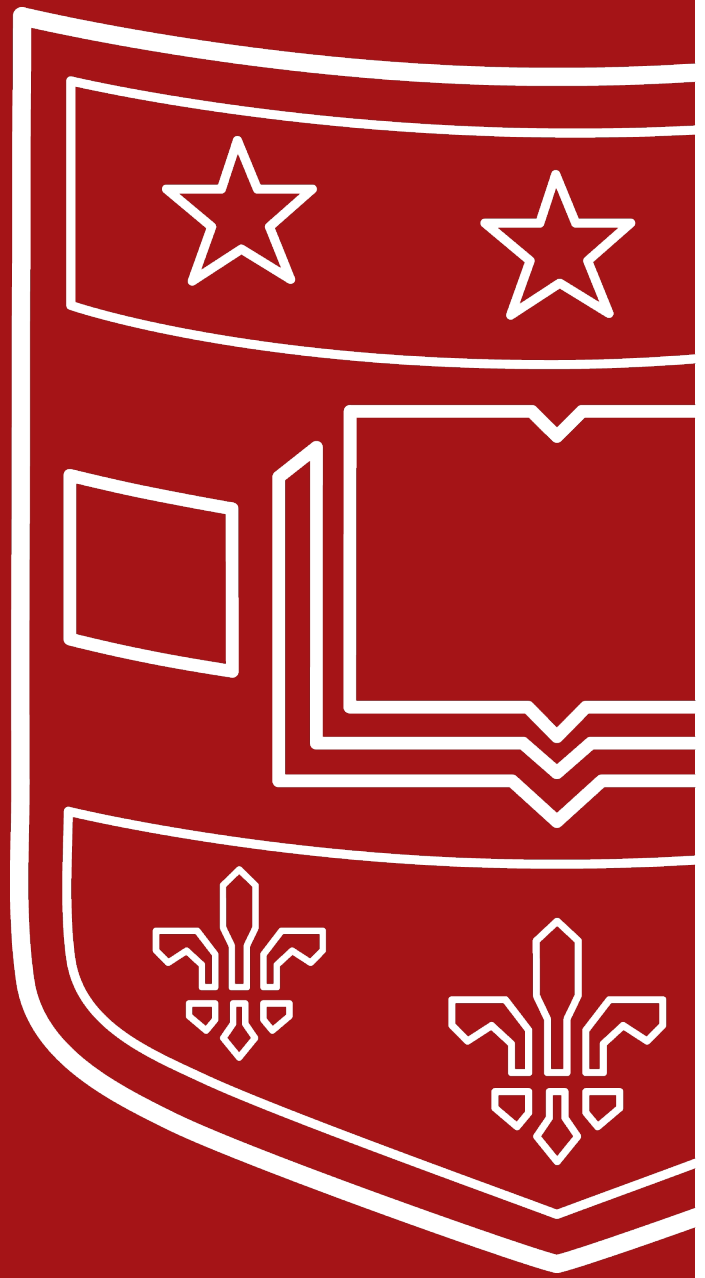
PaLM-E: An Embodied Multimodal Language Model

Yunlai Chen, 04/16/2024

Washington University in St. Louis

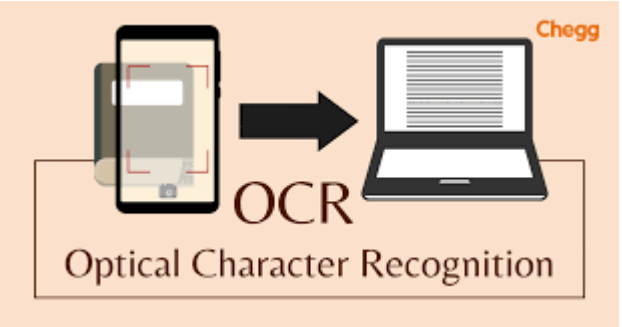


Abstract & Introduction

- Challenge: Real-world application of LLMs faces limitations, like robotics problem.
- The paper introduces embodied language models, integrating continuous sensor modalities from embodied agents with LLMs, aiming for foundational inference in real-world decision-making.
- It explores multi-task training, showing improved performance and data efficiency across robot tasks, visual-language tasks, and standard language tasks.

Background: Related work

- General vision-language modeling.
  (VQA, captioning, OCR, and object detection)



- Actions-output models.

- LLMs in embodied task planning

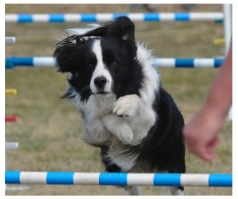
PaLM-E: An Embodied Multimodal Language Model

Mobile Manipulation

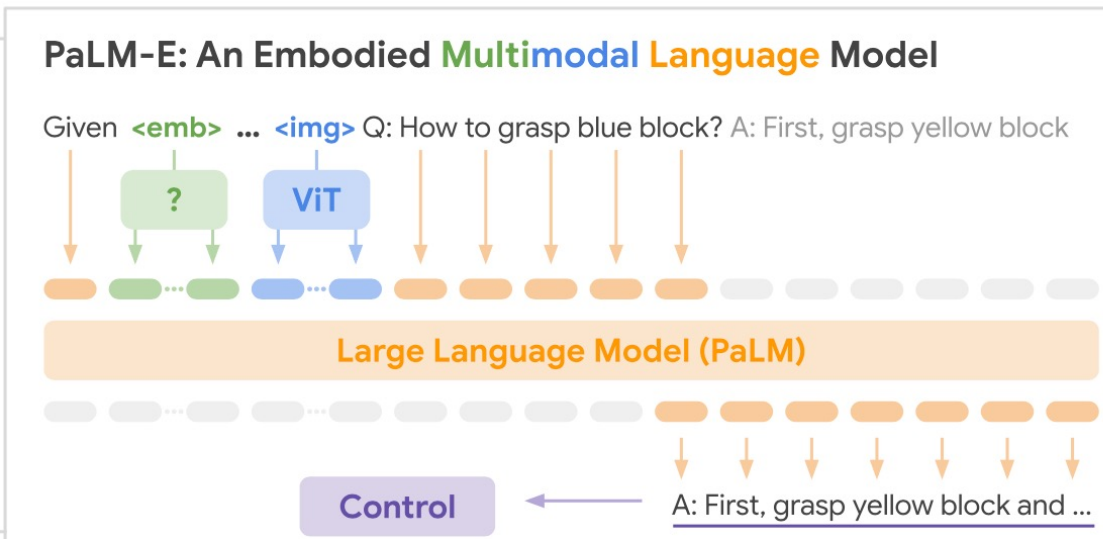
Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Visual Q&A, Captioning ...

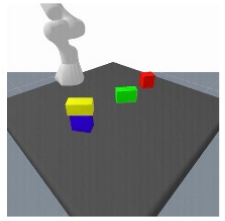
Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.



Describe the following <img>: A dog jumping over a hurdle at a dog show.

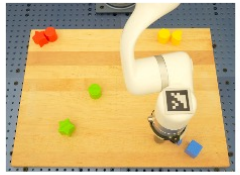


Task and Motion Planning



Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation



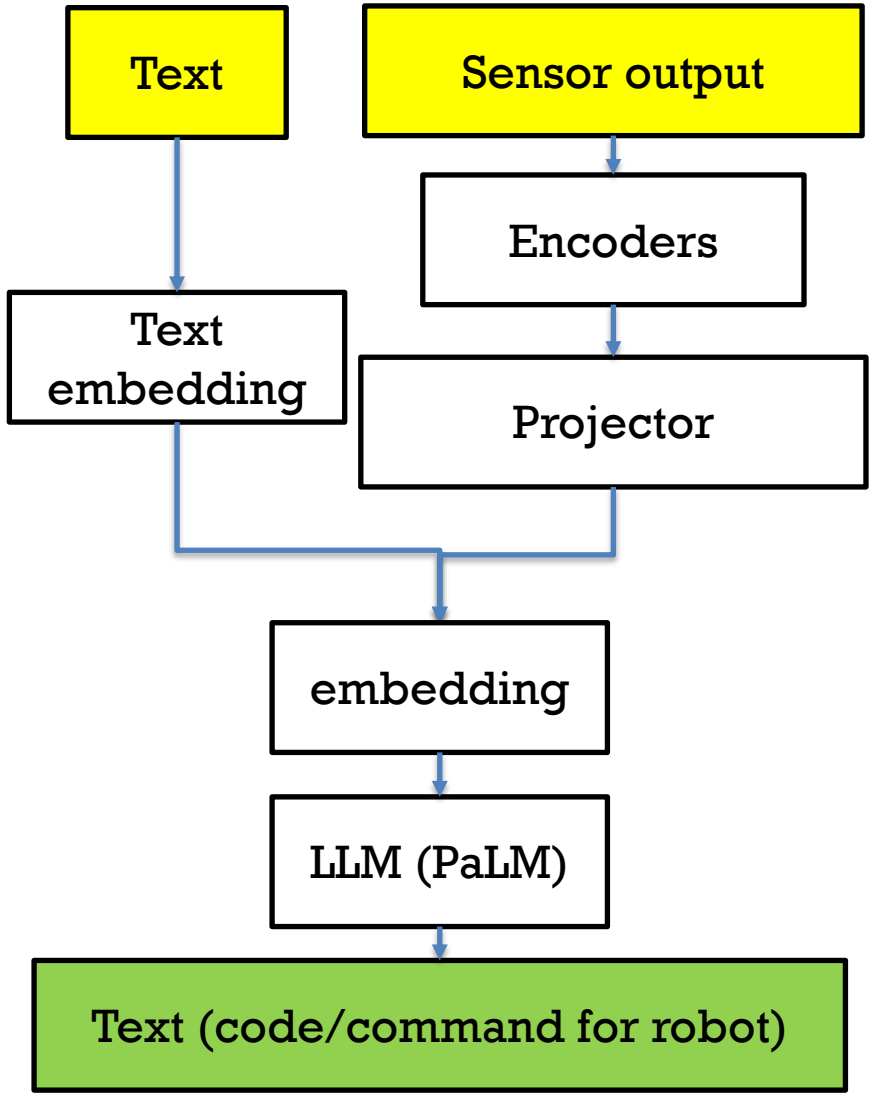
Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

Language Only Tasks

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696.Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

PaLM-E: An Embodied Multimodal Language Model

- Embodied: the model is connected to a robot and has a physical embodiment.
- Multimodal: the model processes inputs from multiple modalities, including text, visual images, and continuous state estimation.
- Language model: the model's output being exclusively textual.



PaLM-E architecture

- Decoder-only LLMs: classic GPT-like structure where the model predicts the next token based on previous tokens

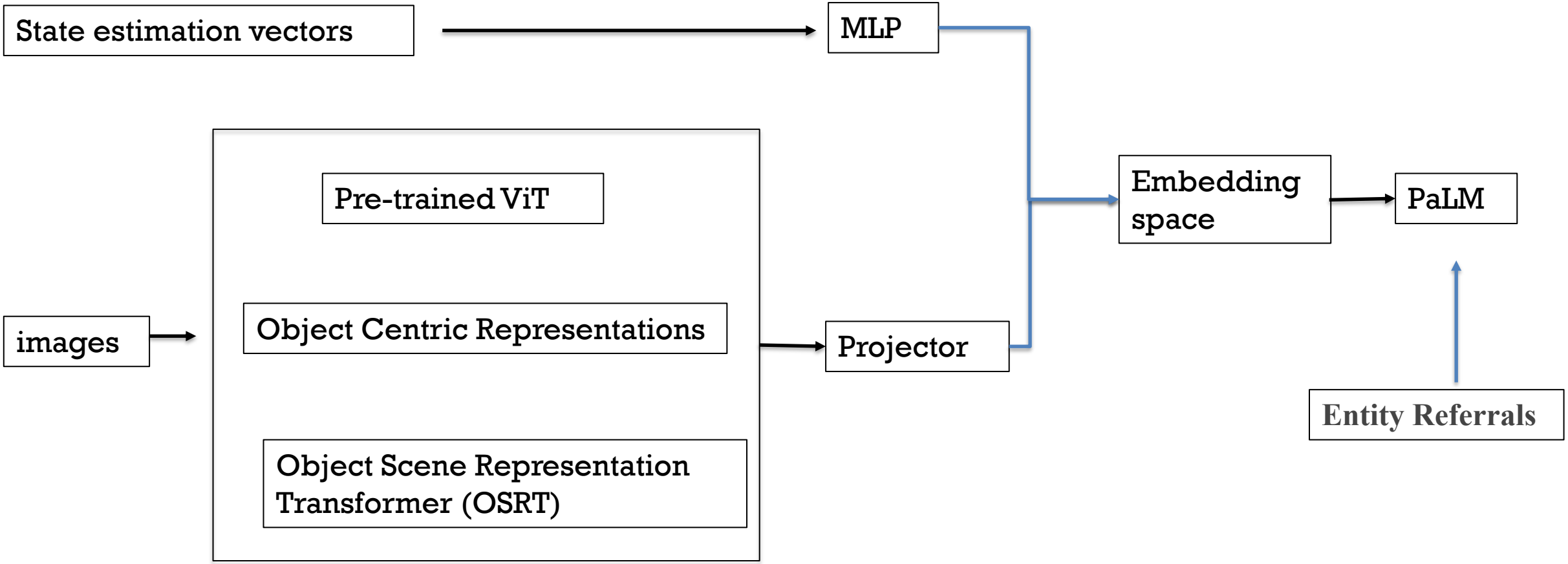
$$p\left(w_{1:L}\right)=\prod_{l=1}^{L} p_{\mathrm{LM}}\left(w_l \mid w_{1:l-1}\right)$$

- Prefix-decoder-only LLMs: Extends the classic model by adding a prefix prompt that doesn't affect loss calculation.

$$p\left(w_{n+1:L} \mid w_{1:n}\right)=\prod_{l=n+1}^{L} p_{\mathrm{LM}}\left(w_l \mid w_{1:l-1}\right)$$

- Text Token Embedding: Maps text tokens to vectors in a predefined embedding space.
- Continuous State Mapping: Translates sensor-detected continuous states into vectors in the same embedding space, requiring multiple vectors for complex states like sounds.

Input & Scene Representations for Different Sensor Modalities



Training

- Dataset

$$D = \left\{\left(I^i_{1:u_i}, w^i_{1:L_t}, n_i\right)\right\}_{i=1}^{N}$$

each instance i includes u_i continuous states observed through sensors, denoted as $I^i_{(1:u_i)}$, and L_i tokens represented by $w^i_{(1:L_i)}$. n_i is the length of the prefix prompt for that data instance.

- Loss function: cross-entropy loss averaged across non-prefix tokens

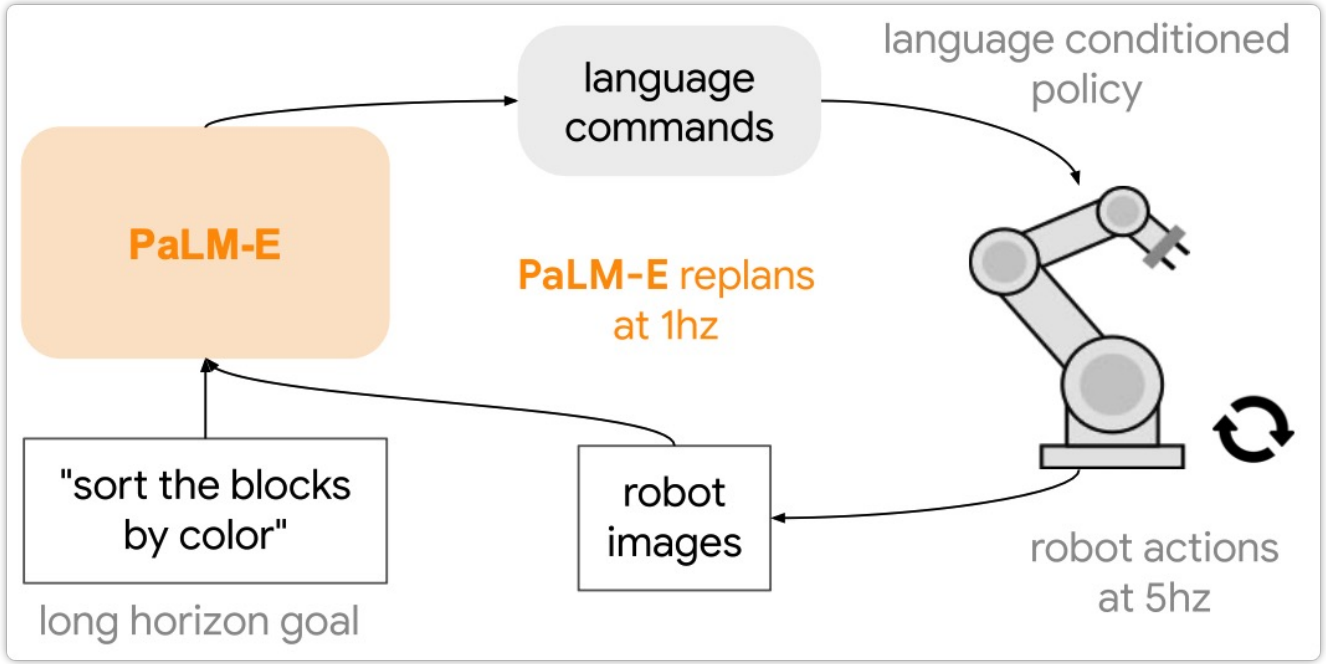
- Variation with Model freezing: fixed LLM, only train input encoders

Experiments & Results: Task and Motion Planning (TAMP) Environment

	ϕ	LLM pre-trained	q_1	q_2	q_3	q_4	p_1	p_2
3 - 5 objects	SayCan (w/ oracle affordances)	✓	-	-	-	-	38.7	33.3
	state	✗	100.0	99.3	98.5	99.8	97.2	95.5
	state	✓(unfrozen)	100.0	98.8	100.0	97.6	97.7	95.3
	state	✓	100.0	98.4	99.7	98.5	97.6	96.0
	state (w/o entity referrals)	✓	100.0	98.8	97.5	98.1	94.6	90.3
	ViT + TL (obj. centric)	✓	99.6	98.7	98.4	96.8	9.2	94.5
	ViT + TL (global)	✓	-	60.7	90.8	94.3	70.7	69.2
	ViT-4B (global)	✓	-	98.2	99.4	99.0	96.0	93.4
	ViT-4B generalist	✓	-	97.1	100.0	98.9	97.5	95.2
	OSRT	✓	99.6	99.1	100.0	98.8	98.1	95.7
6 objects	state	✗	20.4	39.2	71.4	85.2	56.5	34.3
	state	✓	100.0	98.5	94.0	89.3	95.3	81.4
	state (w/o entity referrals)	✓	77.7	83.7	93.6	91.0	81.2	57.1
8 objects	state	✗	18.4	27.1	38.1	87.5	24.6	6.7
	state	✓	100.0	98.3	95.3	89.8	91.3	89.3
	state (w/o entity referrals)	✓	60.0	67.1	94.1	81.2	49.3	49.3
6 objects + OOD tasks	state (8B LLM)	✗	-	0	0	72.0	0	0
	state (8B LLM)	✓	-	49.3	89.8	68.5	28.2	15.7
	state (62B LLM)	✓	-	48.7	92.5	88.1	40.0	30.0

- This experiment used full TAMP training data without other tasks' data.
- Performance was similar when testing with 3-5 objects.
- Increasing objects to 6-8 during testing showed significant improvement with pretrained LLMs.
- For out-of-distribution objects like cups and toy frogs, models without pretrained LLMs failed, whereas pretrained LLMs and increasing model size from 8B to 62B improved performance notably.

Experiments & Results: Interactive Language Table



Input: sort blocks by colors into corners

Output:
"push the red star to the top left corner
"push the red circle to the red star
"push the blue triangle to the blue cube
...
"push the green star to the bottom left corner

long-horizon tasks:
These tasks require models to iteratively generate strategies, execute actions, and assess environmental changes until the objectives are achieved, involving multiple cycles of planning and interaction.

Demo:
https://palm-e.github.io/

Image courtesy: https://mingchao.wang/t5CaYAJ3/#61-tamp

Experiments & Results: Mobile Manipulation Environment

Baselines				Failure det.	Affordance
PaLI (Zero-shot) (Chen et al., 2022)				0.73	0.62
CLIP-FT (Xiao et al., 2022)				0.65	-
CLIP-FT-hindsight (Xiao et al., 2022)				0.89	-
QT-OPT (Kalashnikov et al., 2018)				-	0.63
PaLM-E-12B trained on	from scratch	LLM+ViT pretrain	LLM frozen		
Single robot	✓	✗	n/a	0.54	0.46
Single robot	✗	✓	✓	**0.91**	0.78
Full mixture	✗	✓	✓	**0.91**	0.87
Full mixture	✗	✓	✗	0.77	**0.91**

- Affordance Prediction: Involves predicting whether an action is feasible with a given object/environment, such as determining if a green block under a red block can be moved without first removing the red block.
- Failure Detection: After executing an action, robots must assess whether the action was successful.
- Long-horizon Planning: https://palm-e.github.io/

Experiments & Results: General Visual-Language Tasks

Model	VQAv2 test-dev	VQAv2 test-std	OK-VQA val	COCO Karpathy test
Generalist (one model)				
PaLM-E-12B	76.2	-	55.5	135.0
PaLM-E-562B	80.0	-	**66.1**	138.7
Task-specific finetuned models				
Flamingo (Alayrac et al., 2022)	82.0	82.1	57.8†	138.1
PaLI (Chen et al., 2022)	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0
Generalist (one model), with frozen LLM				
(Tsimpoukelli et al., 2021)	48.4	-	-	-
PaLM-E-12B frozen	70.3	-	51.5	128.0

Although not the main focus of the paper, results for general visual language tasks were reported, including OK-VQA, VQA v2, and COCO Captions.

PaLM-E also achieved top performance on VQA v2 with a frozen LLM, demonstrating its competitiveness as a visual language model and its efficacy in specific reasoning for robotic tasks.

Summary of Experiments & Discussion

- Generalist vs. Specialist Models: PaLM-E's training across different tasks and datasets significantly enhances performance, showcasing its superior transfer capabilities.
- Data Efficiency: Despite the scarcity of robotics data, PaLM-E demonstrates effective learning from limited examples, enhancing its practicality in real-world applications.
- Retention of Language Capabilities: The model retains linguistic skills through strategies like freezing the LLM during multimodal training, or training the entire model end-to-end to minimize catastrophic forgetting.
- Future Directions: Suggestions include leveraging large-scale visual data and continuing to enhance the model's architecture for better performance in embodied reasoning tasks.

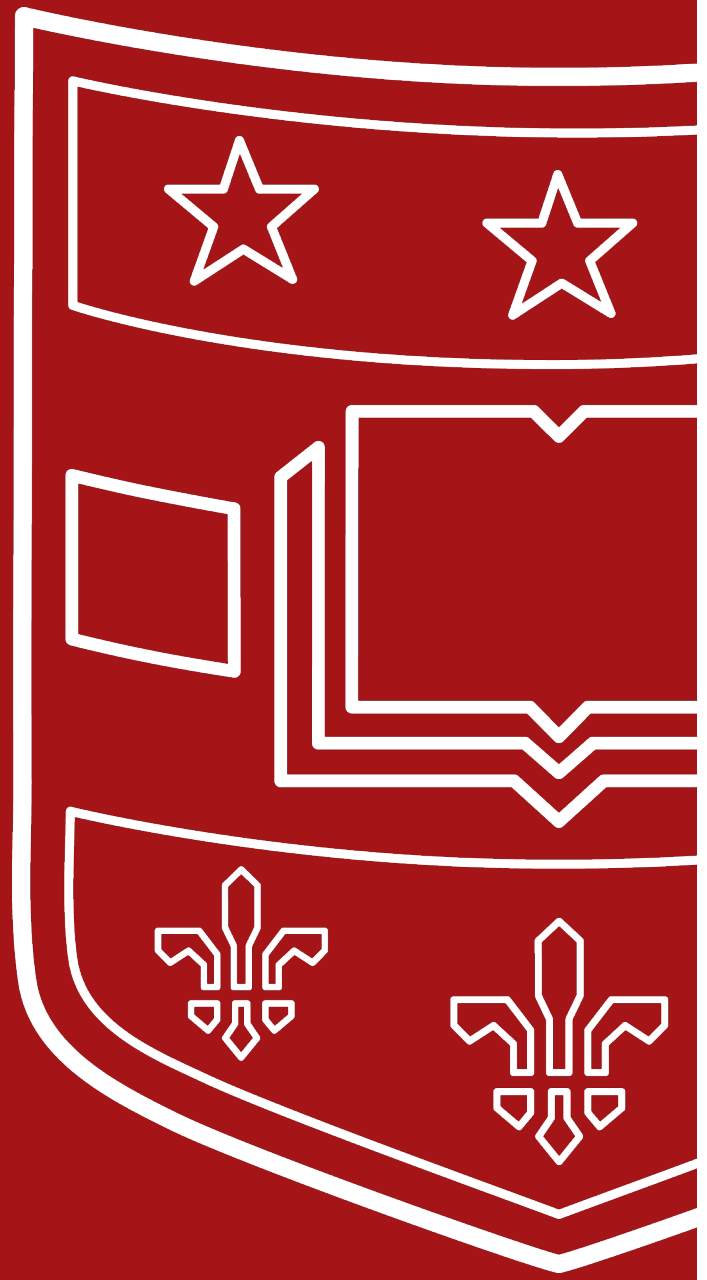
Contribution of this work

- Integrated embodied data into multimodal language models to train a versatile, multi-decision agent.
- Confirmed the feasibility of training image-language models for embodied reasoning, despite limitations in zero-shot models.
- Developed innovative architectural and training strategies for these models.
- Demonstrated that PaLM-E performs well in both specialized reasoning and standard image-text tasks.
- Showed that increasing model size helps mitigate catastrophic forgetting during multimodal fine-tuning.

When Large Language Models Meet Personalization: Perspectives of Challenges and Opportunities

Yunlai Chen, 04/16/2024

Washington University in St. Louis



Introduction

Personalization in AI

- Tailoring AI to individual user preferences and behaviors
- Enhances user experience and engagement
- Examples: Customized recommendations, adaptive learning

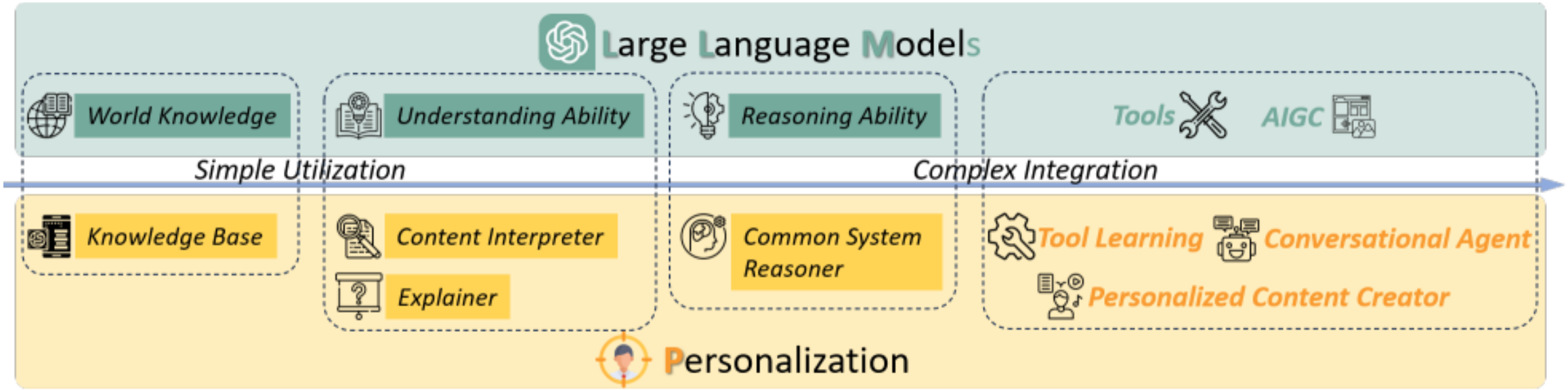
Significance of Integrating LLMs with Personalization

- Merges the cognitive depth of LLMs with user-specific customization
- Potential benefits: Increased satisfaction, higher engagement, better privacy

Purpose of the Paper

- Explores challenges and opportunities of LLMs with personalization
- Aims to identify innovative pathways for personalized AI applications

Background and context



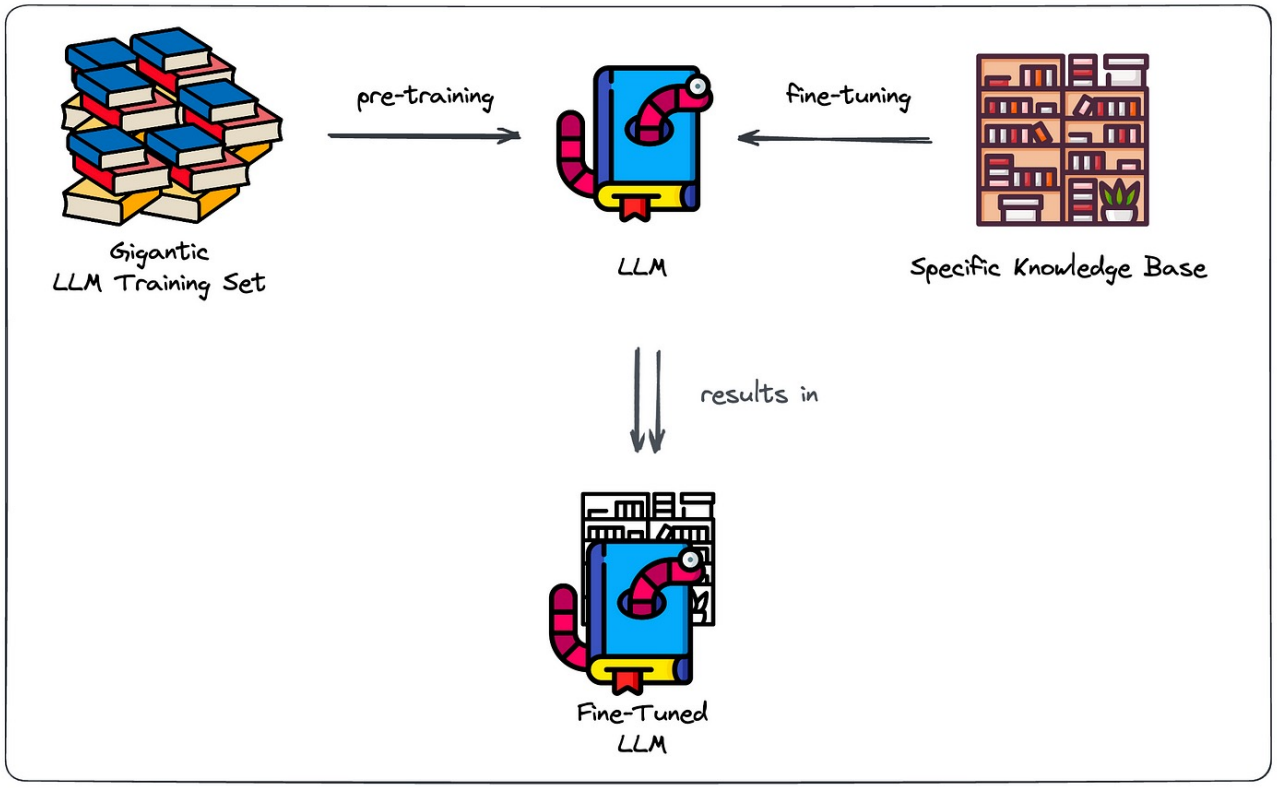
LLMs as Knowledge base

- LLMs can serve as dynamic, self-updating knowledge bases capable of storing and retrieving vast amounts of factual and commonsense knowledge.
- Improves accuracy and explainability through semantic relationships.

- Example:

Semantic Representation: CKE, DKN, SHINE.

Path-Based Enhancement: Hete-MF, SemRec.



LLM as content interpreter

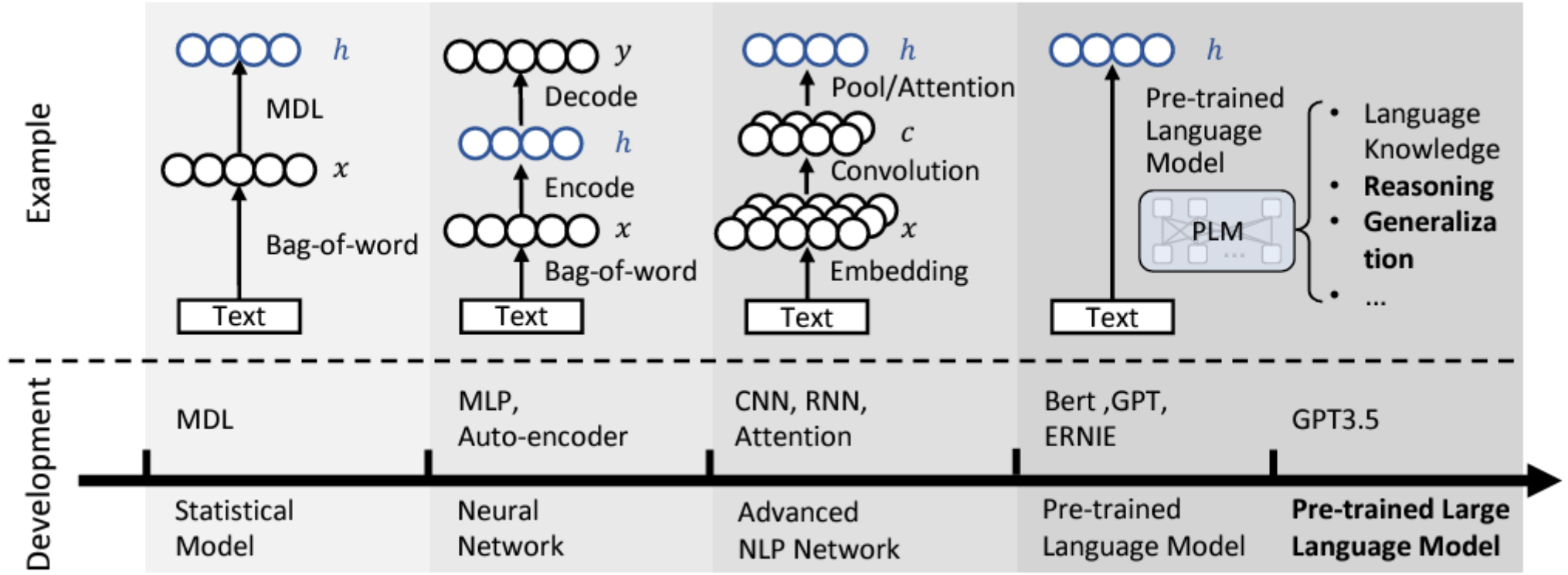


Fig. 2. The development of content interpreter in recommendation.

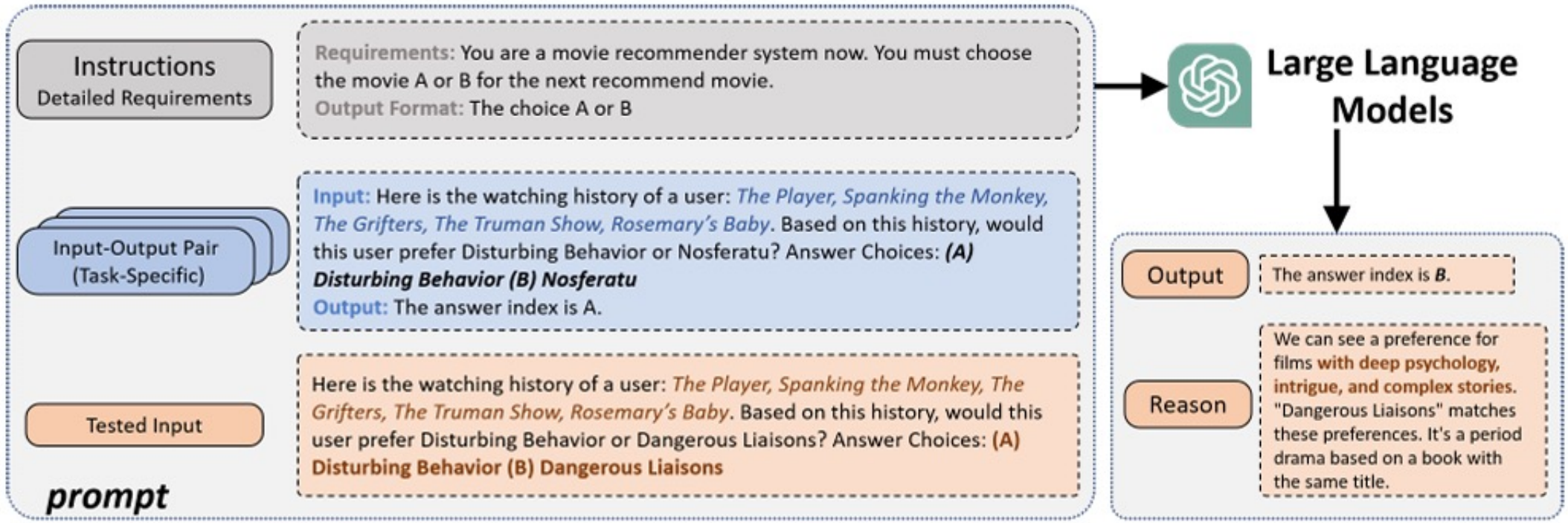
Challenges:
Misalignment of Objectives.
Inference Latency.

LLM as Explainer

- Role of LLMs as Explainers:
  - Purpose: Enhance transparency and trust in AI decisions.
  - Impact: Bridge complex AI operations and user understanding.
- Key Benefits:
  - Transparency: Clarifies decision-making processes.
  - User Trust: Increases confidence in AI recommendations.
  - Compliance: Meets industry standards for explainability.
- Challenge:
  - Opacity/Sincerity/Data Biases/Inconsistency

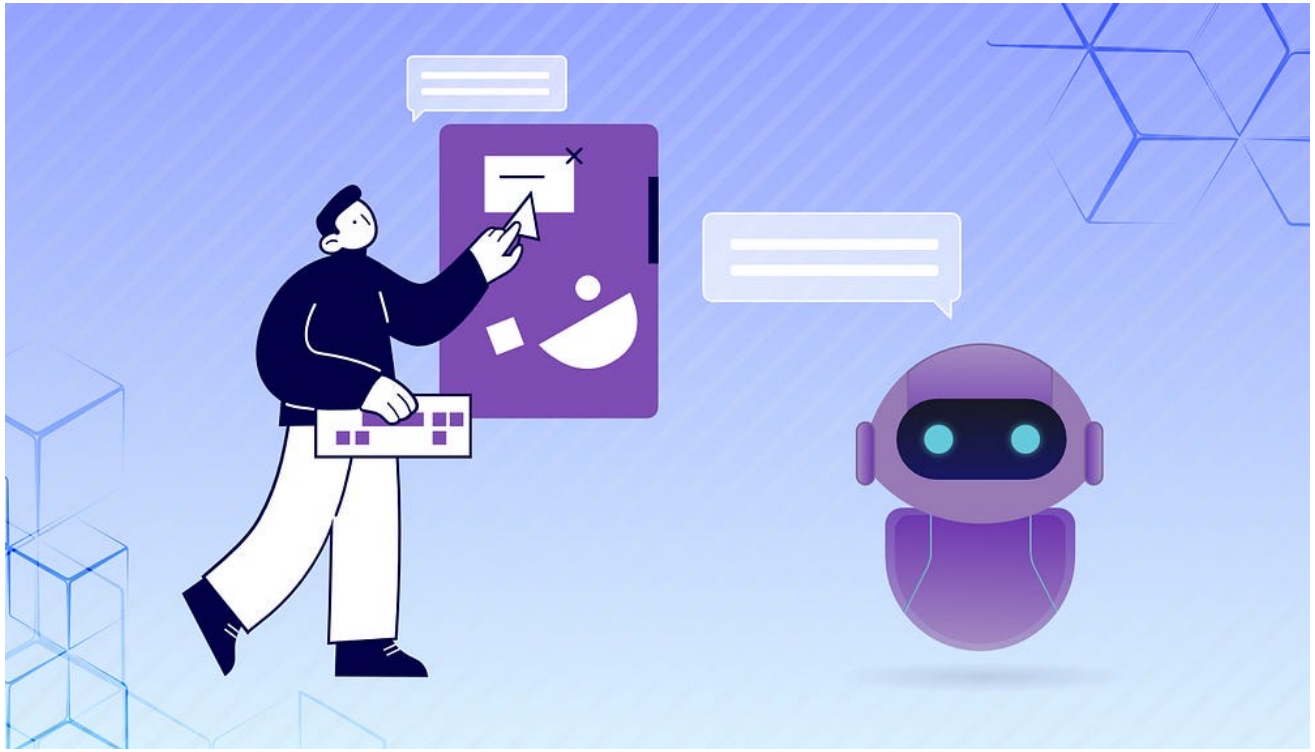
LLM as common system reasoner

- Applications in Recommender Systems:
  - In-context Learning: Enables zero-shot/few-shot learning for direct recommendations without specific model tuning



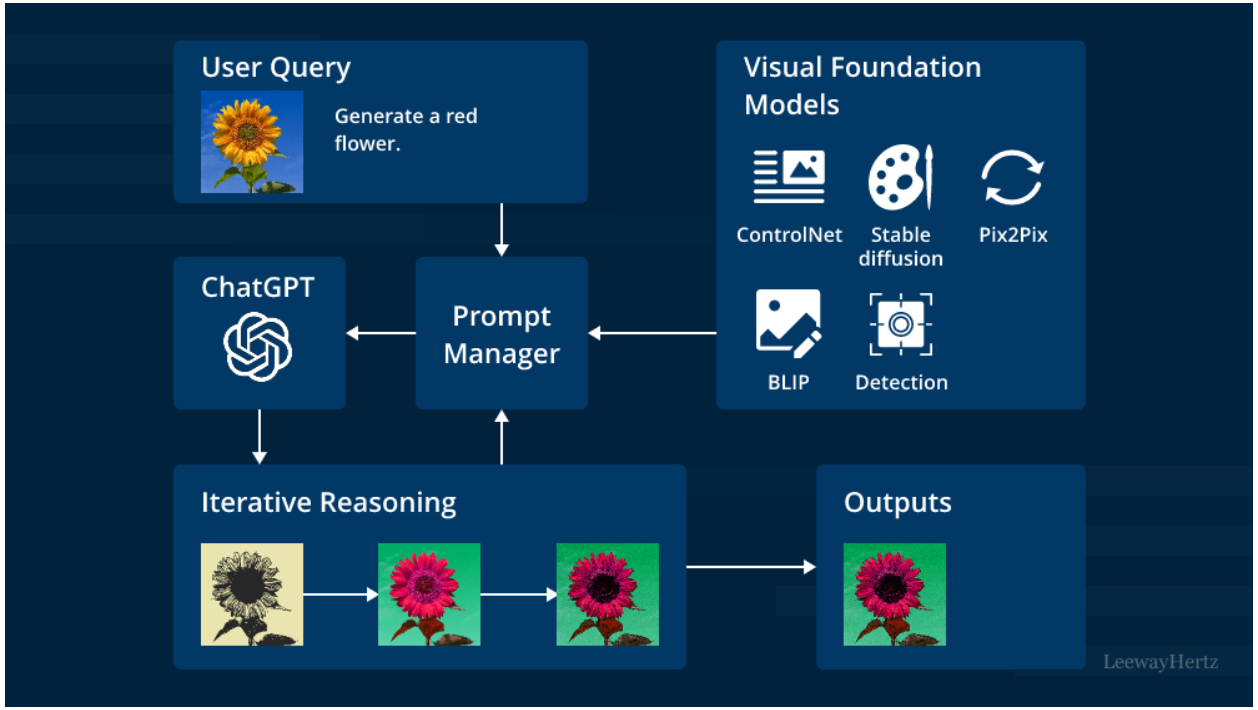
LLM as conversational agent

- Conversational Recommender Systems (CRS)
  - Purpose: Engage users through dialogue to uncover interests and provide personalized recommendations.
  - Advantages: Real-time understanding of user intents and adaptive recommendations based on feedback.



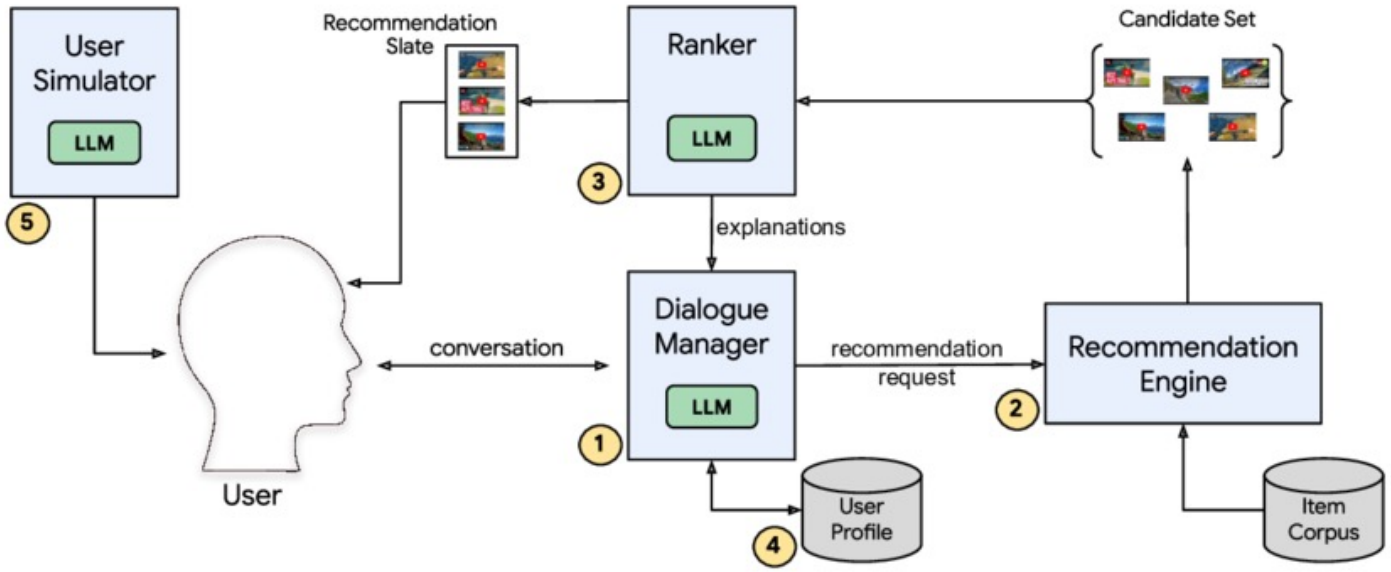
Tool Learning with LLMs in Recommendation Systems

- Introduction to Tool Learning
  - Definition: Combining specialized tools with foundational models to enhance task-solving capabilities.
  - Tool-augmented learning: Tools assist in improving task quality and accuracy.
  - Tool-oriented learning: Focus on training models to optimize the use of tools.
- Applications and examples of LLMs in Tool Learning
  - Capabilities: LLMs decompose complex tasks into manageable sub-tasks, turning them into executable instructions.
  - Visual ChatGPT: Combines visual models with LangChain for visual tasks.



Enhancing Personalization with LLM-Based Tools

- Application:
  - Search Engines: Enhance accuracy and reduce memory load (e.g., BlenderBot 3, LaMDA).
  - Recommendation Engines: Use retrieval and reranking for precise item suggestions (e.g., Chat-REC, RecLLM).
  - Databases: Tackle new item introductions and cold-start issues.



RecLLM

Open Challenges in Industrial Personalization with LLMs

- Scaling Computational Resources
- Achieving Efficient Response Times
- Laborious Data Collection
- Long Text Modeling
- Interpretability and Explainability:
- Evaluation Metrics

Contribution of this paper

- Personalization Systems: The paper reviews current personalization techniques, including recommender systems, personalized assistance, and search, and how LLMs can enhance these areas.
- Emergent Abilities: It explores the emergent abilities of LLMs, such as in-context learning, instruction following, and step-by-step reasoning, which can be leveraged for more sophisticated personalization.
- Challenges and Opportunities: The authors identify challenges in integrating LLMs into personalization systems and propose potential solutions to address them. They also discuss the ethical and privacy concerns associated with LLMs in personalization.

Thanks for your attention

Q&A