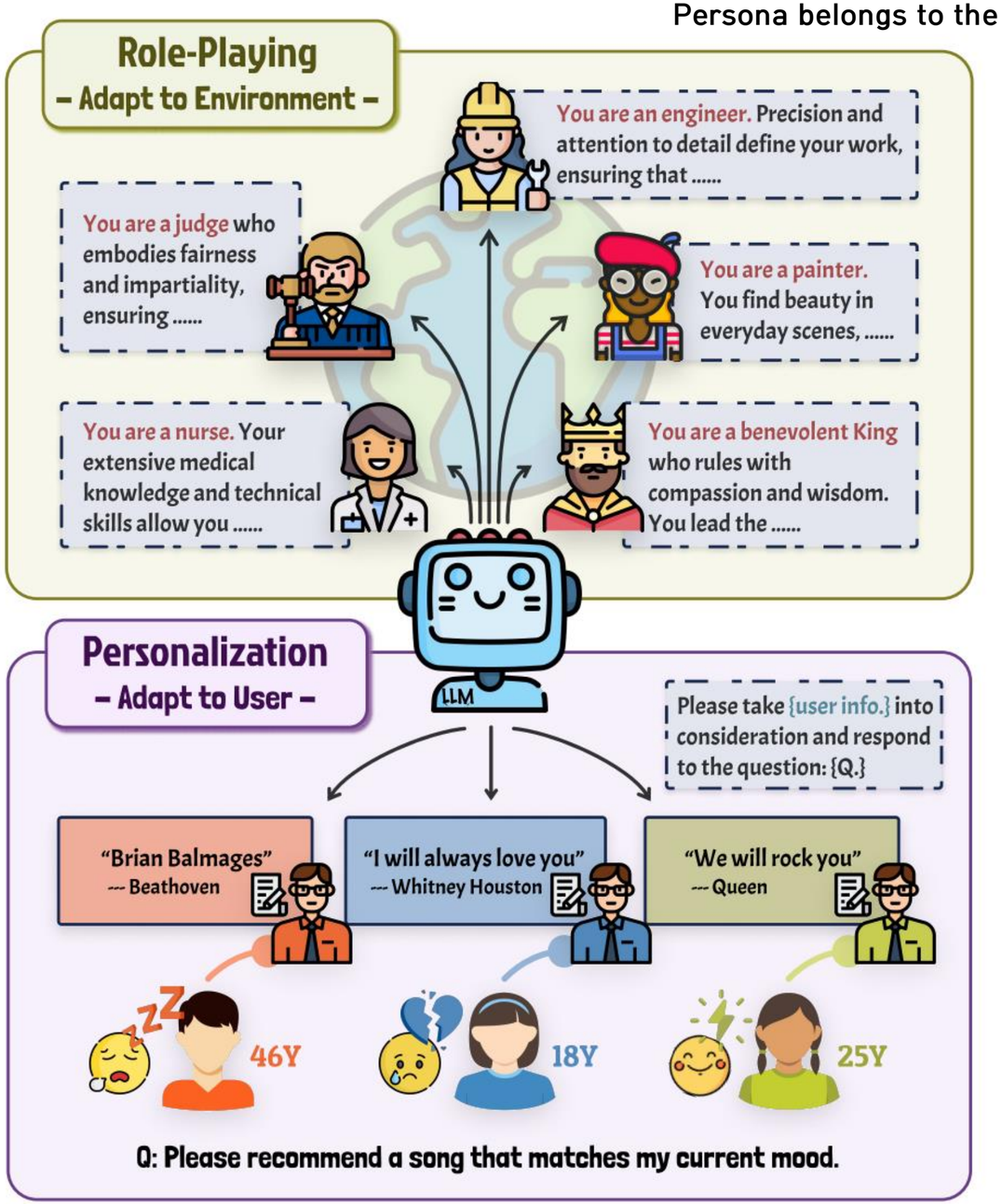
Two Tales of Persona in LLMs

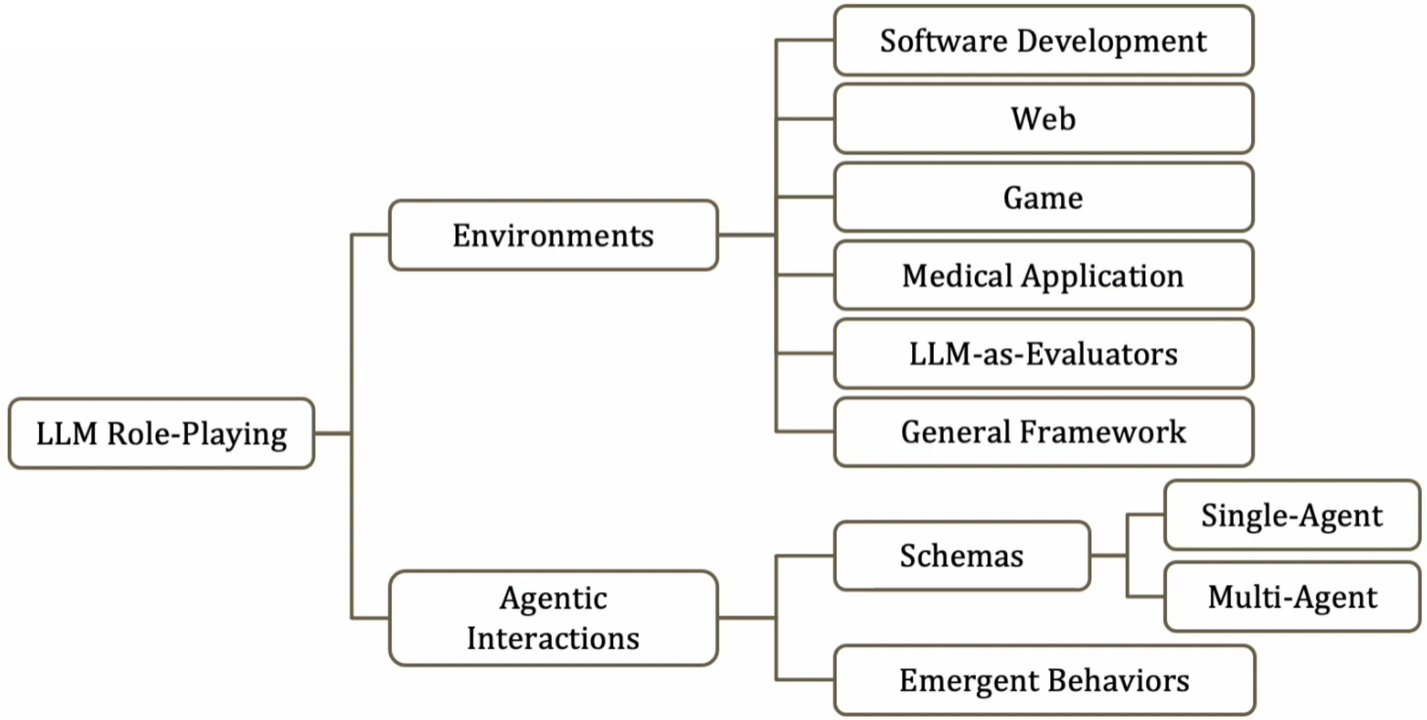
Role-Playing and Personalization

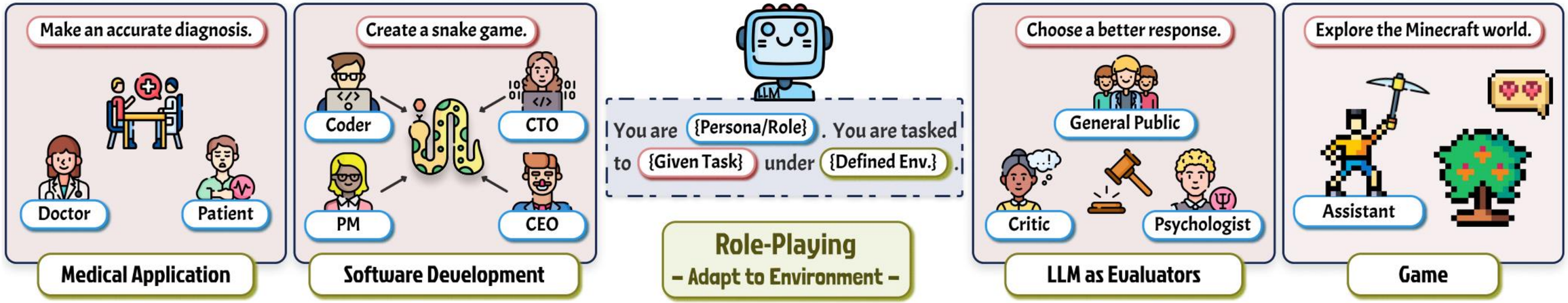
- LLM Role-Playing：LLMs act according to assigned personas (i.e., roles) under a defined environment. For example, given role names with descriptions, LLMs role-play in a social simulation game.
  - How LLMs adapt to defined environments?
- LLM Personalization：LLMs consider user personas(e.g., background information, historical behaviors) to generate tailored responses to the same question.
  - How LLMs adapt to distinct user?



LLM Role-Playing

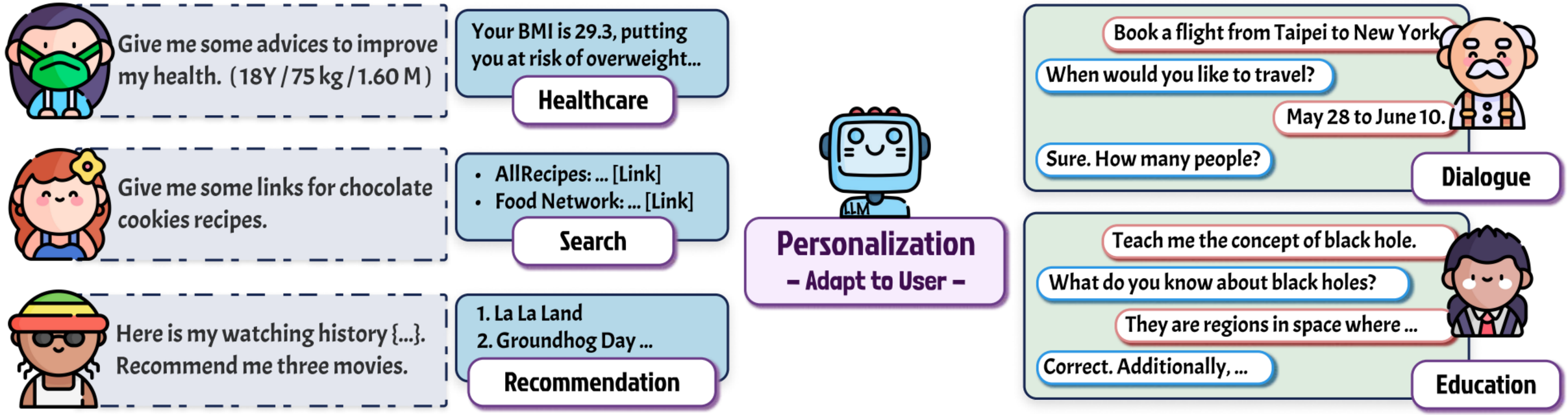
- How LLMs adapt to defined environments, the predominant approach is **LLM-based agents with role-playing.**
- Within environments, LLMs could operate and interact given certain rules and receive feedback.
- Schemas : Single Agent (Game) / Multi-Agent (Software Development & Medical Application) (Cooperative & Adversarial)
- Emergent Behaviors : Volunteer / Conformity / Destructive





LLM Personalization

- Prominent approaches for aligning LLMs to user intents typically leverage **reinforcement learning from human feedback (RLHF)**.
- To enhance individual experience and preference, personalized LLMs consider **user personas(e.g., individual information, historical behaviors)** and cater to customized needs.
- Various personalized tasks with associated methods for achieving personalization. (ex. Dialogue – Policy model, Retrieval + prompting / Recommendation – Prompting, Retrieval, Embeddings, Fine-tuning)



논문 발표 순서

- 1. LaMP: When Large Language Models Meet Personalization (ACL 2024) — User의 페르소나
- 2. User Embedding Model for Personalized Language Prompting (ACL, Personalization of Generative AI Systems Workshop 2024) — User의 페르소나
- 3. LLMs + Persona-Plug = Personalized LLMs (ACL 2025) — User의 페르소나
- 4. SynthesizeMe! Inducing Persona-Guided Prompts for Personalized Reward Models in LLMs (ACL 2025) — User의 페르소나
- 5. Persona Vectors: Monitoring and Controlling Character Traits in Language Models (Anthropic, 2025) — LLM의 페르소나