Reinforcement Learning for Language Model Training

Polina Tsvilodub

LLMs: Outlook

RL4 LMT

Limitations & social implications of LLMs

Summaries

- McCoy et al. (2023):
  - LLMs' performance is sensitive to task probability, input probability and output probability
- Jo & Gebru (2020):
  - when collecting training data for systems like LLMs, the ML community should pay more attention to systematicity in quality of data collection
- Hendricks et al. (2021):
  - in order to test alignment of LLMs to human values, datasets like ETHICS are developed (for testing predictions of various ethical judgements) — LLMs have far from perfect alignment
- Santurkar et al. (2023):
  - LLMs are biased towards reflecting opinions of certain subgroups in the US population, and are inconsistent across topics — general population is not reflected
- Shah et al. (2022):
  - even correctly trained RL systems might misgeneralize learned behavior (and the pursued goals) in test situations which differ from training environments
- Pathak et al. (2017):
  - including an 'internal' curiosity model for learning about the environment features which are relevant to the agent improves its generalisation

LLMs as agents

LLMs as building blocks

- AutoGPT:
  - based on GPT, autonomously generates “thoughts” to achieve a user-specified goal
    - including continuous execution mode
  - internet access for searches and information gathering
  - memory management
  - GPT-4 instances for text generation
  - file storage and summarization with GPT-3.5
  - extensibility with Plugins
    - TTS, code execution, emails, trading…

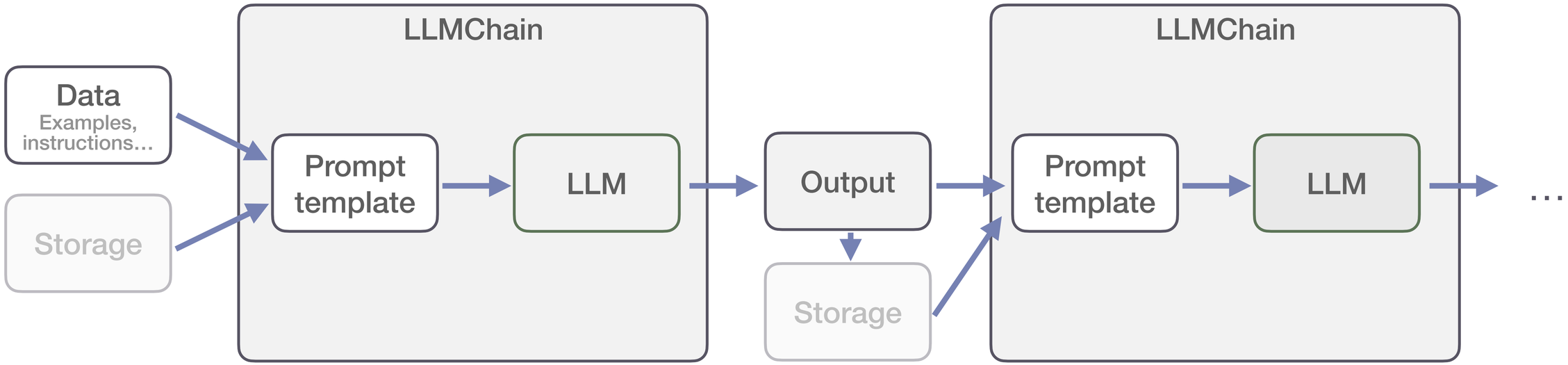
DO NOT RUN ON YOUR MAIN MACHINE!

AutoGPT, BabyAGI, JARVIS

LangChain Chains

$10 million dollar baby

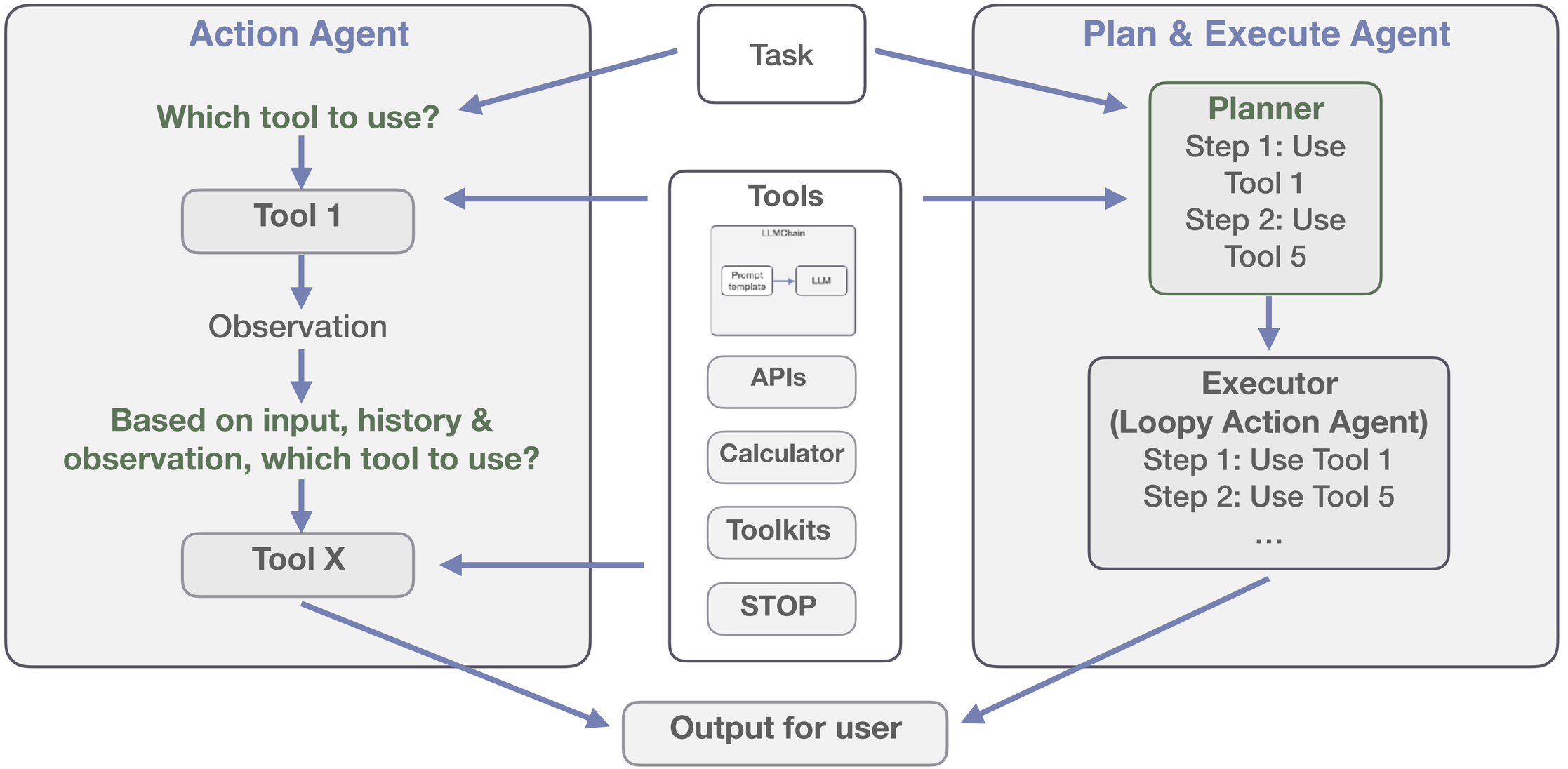
"a framework for developing applications powered by **language models**" can also be *data-aware* and *agentic*



source

LangChain Agents

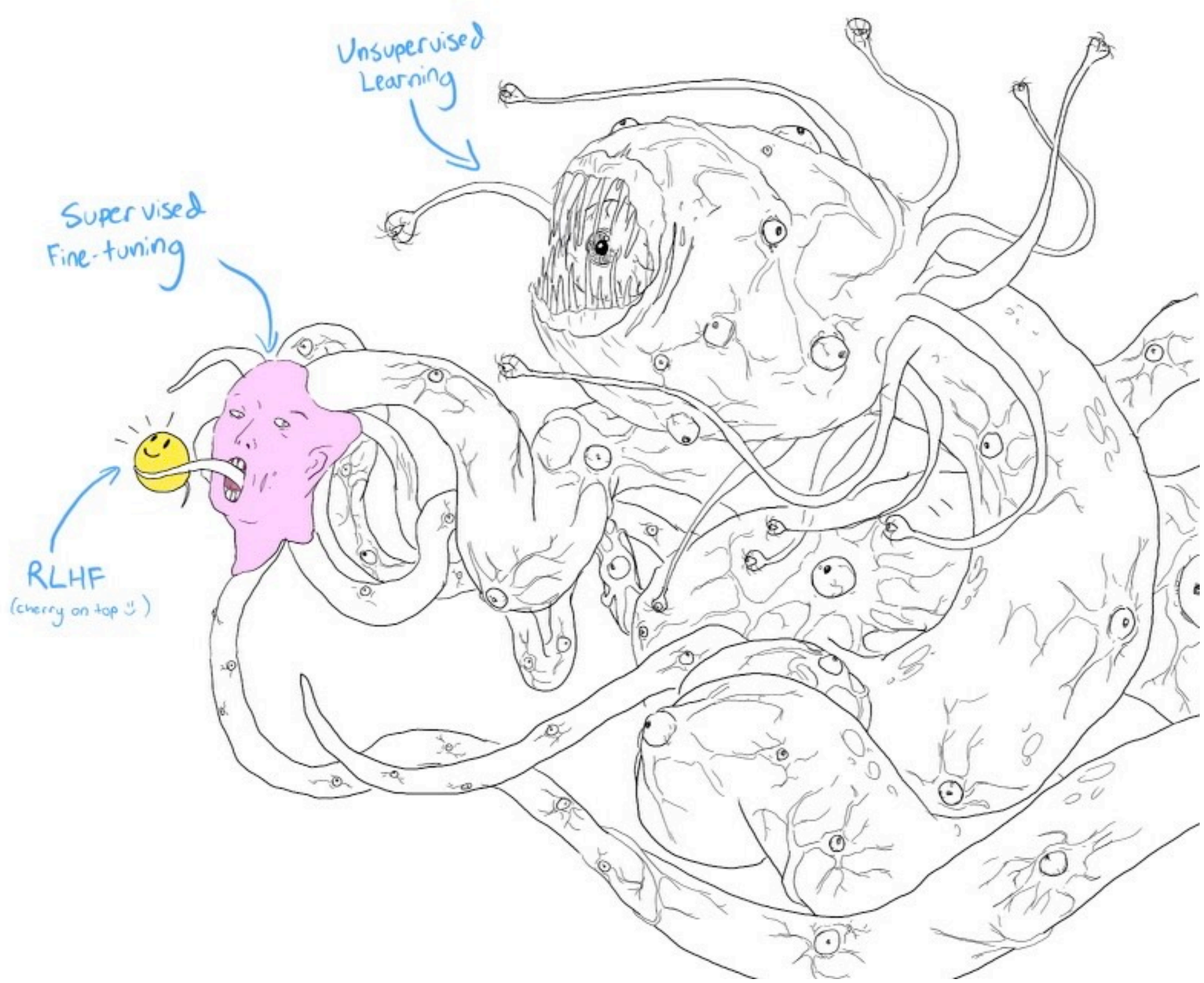
Implementing an unknown chain defined based on input



source

How to think about LLMs?

Shoggoth



source here

How to think about LLMs?

Classifiers, agents, simulators, …

[Opinions ahead]

- LLMs can be prompted into different **personas** (Wolf et al., 2023)
  - personas can be seen as different mixtures of traits
  - different personas might facilitate jailbreaking
  - RL fine-tuning might facilitate adversarial prompting
- **Waluigi effect**: “After you train an LLM to satisfy a desirable property P, then it's easier to elicit the chatbot into satisfying the exact opposite of property P.”
- LLMs are different from other model types
  - **simulators**: “optimized to generate *realistic* models of a system”
  - **simulacrum**: particular instance generated by simulator

Wolf et al. (2023), Waluigi blogpost, simulators blogpost, image

LLMs as building blocks

"The key observation is that large language models **encode a wide range of human behavior** represented in their training data. […] With their ability to **generate and decompose action sequences**, large language models have also been used in planning […]."

"[…] we compare GPT-4 to ChatGPT throughout to showcase a giant leap in level of **common sense** learned by GPT-4 compared to its predecessor."

Park et al. (2023), Microsoft (2023)