DS-GA 3001 005 | Lecture 6

Reinforcement Learning

Jeremy Curuksu, PhD
NYU Center for Data Science
jeremy.cur@nyu.edu

March 13, 2024

DS-GA 3001 RL Curriculum

Reinforcement Learning:

- Introduction to Reinforcement Learning
- Multi-armed Bandit
- Dynamic Programming on Markov Decision Process
- Model-free Reinforcement Learning
- Value Function Approximation (Deep RL)
- **Examples of Industrial Applications and Project Q&A**
- Policy Function Approximation (Actor-Critic)
- Planning from a Model of the Environment
- Advanced Topics and Development Platforms

Reinforcement Learning

Last week: Value Function Approximation

- Categories of Functions in Reinforcement Learning
- Approximation of State Update Functions
- Approximation of Value Functions
- Deep Reinforcement Learning

Today: Examples of Industrial Applications

- **A Tour of 10 Awesome Applications of RL**
- **Project Q&A**

Robotics

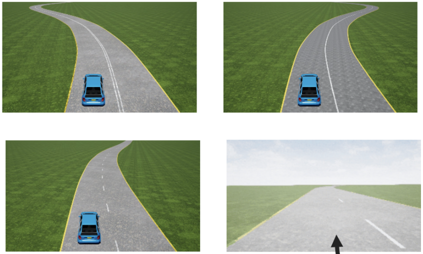
Teach a Robot to ...

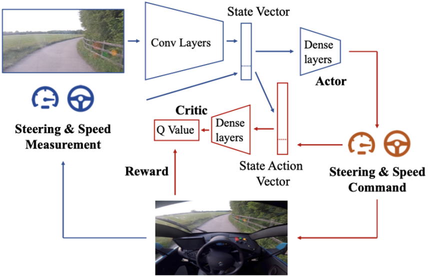
Source: DeepMind (2022)

Autonomous Driving

Learn to Drive Like a Human

- **Goal**: Drive vehicle on a circuit to destination without leaving the road





Environment:

Virtual and physical driving lanes

Reward Function:

- Forward Speed
- Termination upon *infraction of traffic rules* by safety driver

State:

- Pixels of front camera encoded by CNN (single monocular image)
- Vehicle speed and steering angle

Actions:

2 actions: Speed, steering angle

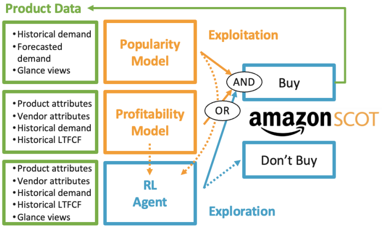
Learn to Drive Like a Human

Source: Wayve (2019)

Amazon Inventory Management

Manage Amazon Retail Inventory

- **Goal**: Select products to show as "shipped and sold by Amazon" on the Amazon.com website to maximize customer experience and profitability



Offline Validation: All auto parts and medical supply products in catalog as of 12/2018. Cash flow and sales in 2019.

Statistics	ASINs selected to buy	ASINs blocked from buying
ASINs count (in MM)	0.16 (2.7%)	5.55 (97.3%)
Cash flow (in MM euros)	152.35	-19.04
Sales (in MM)	28.06 (91.7%)	2.57 (8.3%)

Online A/B testing: Q4 2019, 30M products, 90% treatment, 10% control

EU LAB	Treatment Effect per ASIN per week	Confidence Interval	p-value	Annualized Impact
CP (Euros)	0.0103	[0.002, 0.019]	0.02	€2.45 MM
Sales (Euros)	0.021	[-0.061, 0.103]	0.10	€4.68 MM
Cash flow (Euros)	0.1123	[-0.311,0.536]	0.54	€25.03 MM
Out of stock (bps)	-74	[-100, -50]	0.00	-74 bps

Environment (Contextual Bandit):
Product's profitability and popularity, $\Delta t = 3$ months

Reward Function:

$$\text{Profit} = \sum \begin{pmatrix} \text{short term profit} + \text{long term value} \\ -\text{cost of capital} - \text{asset depreciation} \end{pmatrix}$$

State:
Product-, brand- and vendor-level statistics, historical sales, historical cash flow, glance views

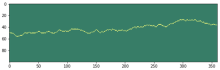
Actions:
1 action: (block buy, buy at least 1 unit)

Seismic Mapping to Identify Natural Oil & Gas Reserves

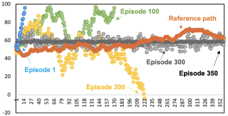
Seismic Pathfinder

- **Goal**: Screen cross-sections under the earth surface to identify the nature and geometry of individual seismic layers, to reduce exploration costs

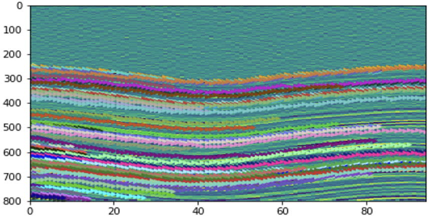
Pathfinder for 1 path:



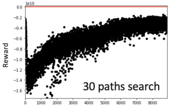
Sample of paths explored by RL agent:

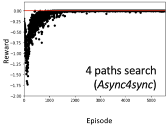


1 cross-sectional (cylinder) map:



Accumulated reward:





Environment:
Underground cylinder cross-sections

Reward Function:

$$-\beta_1 \sum \left|z_t^i - z_{ref}^i\right| - \beta_2 \sum \left|z_{end}^i - z_0^i\right|$$

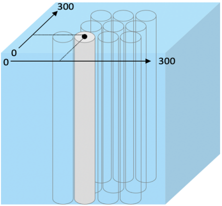
State:
Indirect seismic measurements at z_t and $(z-3, z-2, z-1, z, z+1, z+2,\ z+3)_{t+1}$

Actions:
k actions: (up, down) for each path

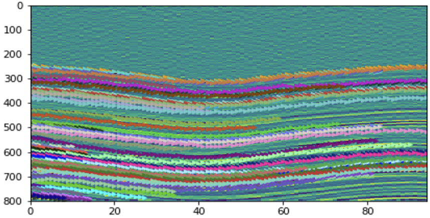
Seismic Pathfinder

- **Goal**: Screen cross-sections under the earth surface to identify the nature and geometry of individual seismic layers, to reduce exploration costs

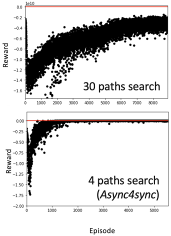
3D seismic cube:



1 cross-sectional (cylinder) map:



Accumulated reward:



Environment:
Underground cylinder cross-sections

Reward Function:

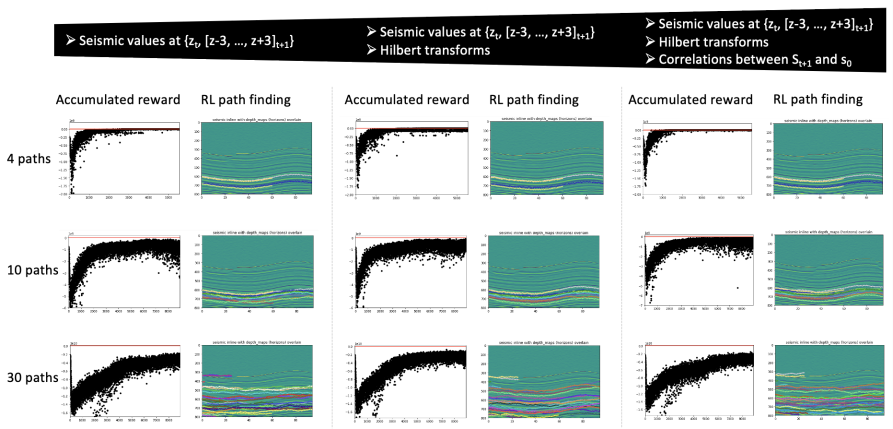
$$-\beta_1 \sum |z_t^i - z_{ref}^i| - \beta_2 \sum |z_{end}^i - z_0^i|$$

State:
Indirect seismic measurements at z_t and $(z-3, z-2, z-1, z, z+1, z+2, \ z+3)_{t+1}$

Actions:
k actions: (up, down) for each path

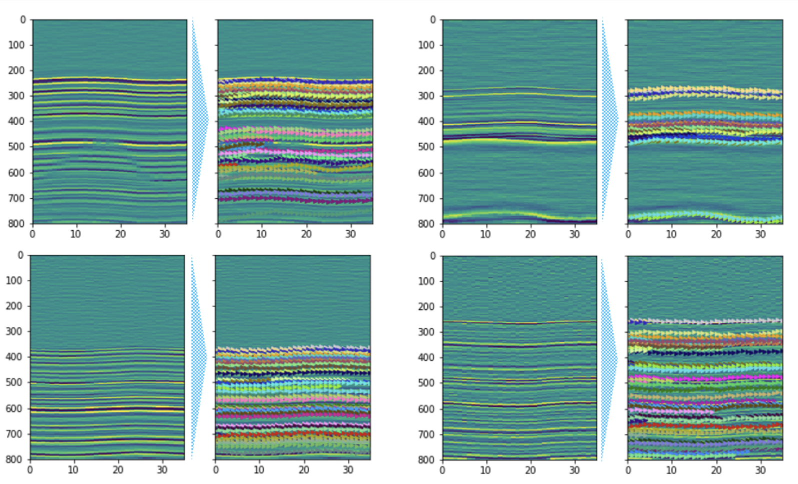
Seismic Pathfinder

- **Analysis of Results**: Benchmarks of synchronous *n*-path search RL on state complexity and number of paths



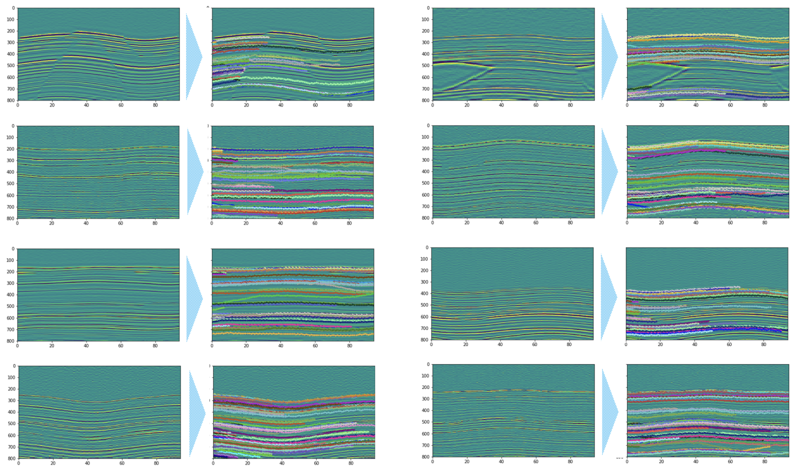
Seismic Pathfinder

- **Analysis of Results**: Generalization of pre-trained Async4sync DRL agent on arbitrary cubes and cylinders with **radius = 35 steps**



Seismic Pathfinder

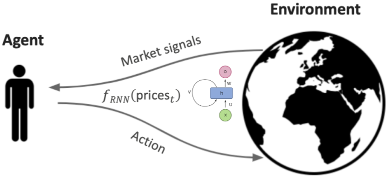
- **Analysis of Results**: Generalization of pre-trained Async4sync DRL agent on arbitrary cubes and cylinders with **radius = 90 steps**



Algorithmic Trading

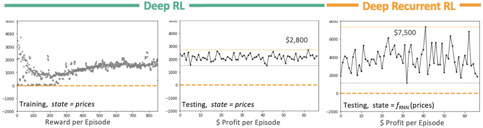
Manage a Stock Portfolio

- **Goal**: Identify trading strategy in portfolio of k stocks to maximize profit



RNN-based RL outperforms lag-based RL at trading stocks

Mean return \$4,900 vs. \$2,200; Maximum return \$7,500 vs. \$2,800



Environment:

7 years of stock prices and news headlines, Δt =1 day

State:

Vector of k stock prices for last n days encoded by RNN

Reward Function:

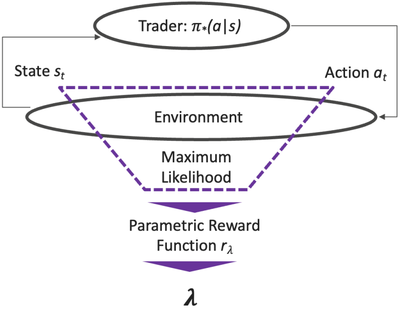
$$r_t = r_t^0 + r_t^{risk} + r_t^{fee} = \underbrace{\sum_k \frac{(\text{prices}_{t+1} - \text{prices}_t)}{\text{prices}_t} x_t}_{\text{Portfolio Value Change}} - \underbrace{\lambda\, \sigma_t^2(r_t^0)}_{\text{Risk Penalty}} - \underbrace{\kappa_t^T x_t}_{\text{Transaction Cost}}$$

Actions:

k actions: (\$buy, \$sell, sit)$_k$

Interpret Financial Trading Behavior

- **Inverse Reinforcement Learning**: Identify trader attributes, such as a level of risk aversion, by observing its behaviors
- **Estimate parameters of a reward function** that fit observed trajectories under a given policy in historical or simulated experience
- **Example:** Compute the risk aversion parameter λ of a successful trader



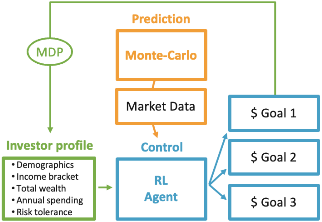
Reverse Engineer Strategy from Trader (= proprietary trading events) using Inverse RL:

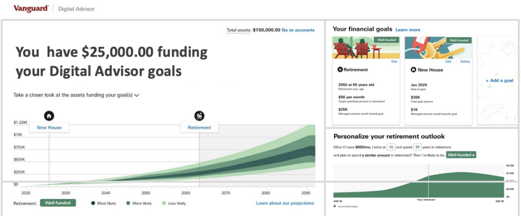
1. Choose a parametric form of reward function
2. Estimate its parameters (MLE) from observed behavior in past trading events until convergence
3. Apply RL under the estimated reward function on an arbitrary stock portfolio, to identify an optimal policy (trading strategy) for this stock portfolio

Asset Allocation and Wealth Management

Manage Long-Term Financial Goals

- **Goal**: Determine optimal asset allocation strategy to meet multiple long-term financial goals, while also being successful in retirement





Environment:

$p(s' \mid s, a)_{stable} + w(\sigma^2)_{MC}$, $\mathbb{E}(p_{MC}, a)_{MC}$, $\Delta t = 1$ year

Observed State:

Investor profile, past $ contribution to each goal

Reward Function:

$$r_t = r_t^{work} + r_t^{retired} = -\beta_1 \sum \left| g_t^i - g_T^i \right| + \beta_2\, p_{MC}$$

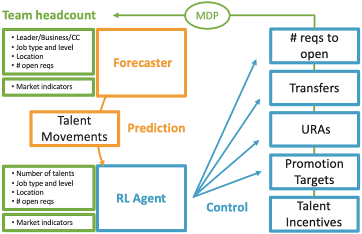
Actions:

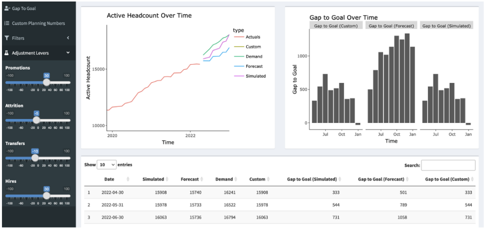
k actions: $ contribution to each goal

Workforce Planning

Manage a Large Talent Workforce

- **Goal**: Pull talent levers to minimize gaps "actual *vs.* target headcount" while also minimizing costs across the World-Wide Amazon workforce





Environment:

$p(s' \mid s, a)_{(internal + external\ factors)}$, $\Delta t = 1$ month

State:

Talent team size, job type, job level, location, number of open jobs, market data, talent movement forecasts

Reward Function:

$$r_t = r_t^{gap} + r_t^{cost} = \beta_1 |h_{EoY} - h_{target}| - \beta_2 c_t$$

Actions:

Jobs to open, Transfers, URA, Promotions, Compensation (...)

Workforce MDP Simulator

- **Model used to define next states**: Forecast monthly talent movement based on historical trends and market indicators

Graph Transformer for Workforce Planning

Jeremy Curuksu
People Experience and Technology
curukj@amazon.com

Jimmy Rigby
People Experience and Technology
rigbjame@amazon.com

Abstract

We present a Graph Transformer deep learning method for workforce planning which can identify potential talent risks and forecast individual monthly talent movements in each segment of the Amazon corporate population. This method outperforms linear methods in current usage at Amazon by 40% in 76% of the segments, and outperforms all state-of-the-art baselines tested by 15% in 61-70% of the segments. Given the dependency Graph in the proposed method can be used to interpret and understand talent forecasts, there is little drawback to using this method compared to using linear trailing rates for workforce planning.

1 Introduction

Workforce planning at Amazon sets year-end headcount targets by financial cost center to meet the company's current and future staffing needs based on business goals and talent movement forecasts (hires, promotions, transfers, attritions). Individual team leaders further plan their workforce needs by individual team, job type, job level and location. Failure to accurately forecast future headcounts and staffing needs often result in delays in productivity and costly resource allocation [1].

Forecasting Amazon talent movements is challenging due to complex spatial and temporal dependencies within the Amazon population, and non-stationarity that result from unusual events such as the Covid pandemic [2]. Amazon is an especially difficult forecasting problem due to its diverse operation workforce and unprecedented size (over 1.6M employees in peak season [2]).

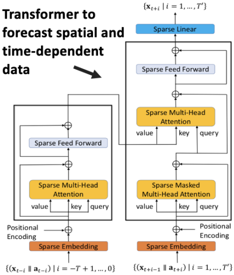
Examples of talent movement dependency at Amazon include talents in similar environments, with similar profiles and job market opportunities, or under similar talent management strategies. Many other factors, including external factors such as large swings in the company's stock value [3], can influence talent flows and thereby create correlated traffic patterns in the Amazon population.

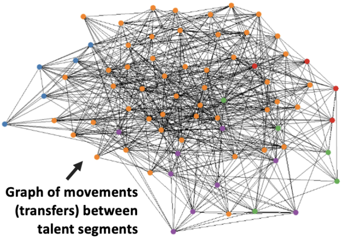
Forecasting spatial and time-dependent data in large, complex traffic networks was recently addressed by estimating a dependency Graph that parsimoniously represents the spatial dependency between different locations in the network (nodes), and using this Graph to sparsify a deep learning Transformer model [4]. Transformers [5, 6] have been shown to significantly outperform RNN and CNN and improve the learning of long-range temporal dependency [5]. By assigning each neuron of the Transformer with a spatial location and using knowledge from the dependency Graph to prune neural connections that are not dependent, the resulting *Graph Transformer* could efficiently capture *both* spatial and temporal dependencies and scale to large traffic networks [4].

In this paper, we use a multivariate Gaussian approximation to find the dependency Graph of talent movements over the different teams of the Amazon corporate population defined by leader, job type, and job level. We use this Graph to derive insights into the overall dynamics of Amazon talent

10th Amazon Machine Learning Conference (AMLC 2022)

Transformer to forecast spatial and time-dependent data





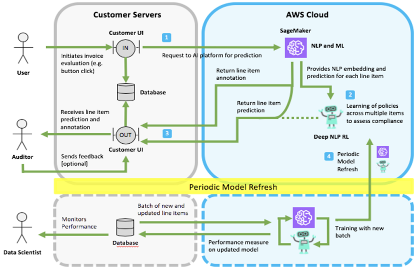
Graph of movements (transfers) between talent segments

Model	MAPE	Nodes improved	MAPE in nodes improved
Graph Transformer	84%	N/A (*self*)	93%
Prophet	95%	70%	108%
s-ARIMA/State Space	87%	61%	107%
Trailing 3-month	112%	76%	133%

Audit Financial Claims with NLP

Audit Claims with Natural Language

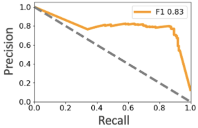
- **Goal**: Recommend compliance level of items in financial claims, to reduce time spent by human contractors, and to reduce errors

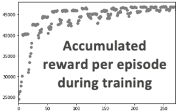


Agent finds 86% of known frauds at 80% precision

Fraud Test Results:

Precision:	80%
Recall:	86%
F1 Score:	0.83
Brier Score:	0.04
AUC Score:	0.95





Environment:

50,000 claims with at least 2 items per claim

Reward Function:

- Compliance category predicted by NLP classifier trained on individual items (regular audit)
- Frauds detected by specialized auditors

State:

- Last *n* items in claims encoded by NLP
- Claim metadata (source, activity code, billed units, …)

Actions:

2 actions: Fraud risk level and compliance category

Audit Claims with Natural Language

- **Post-processing to interpret RL results**: Clustering in NLP space of items at risk can help auditors identify patterns of frauds more quickly

Rearrange non-compliant items per date/source/claim

Source	Claim ID	Date	Description
Mr. Z	61620	10/11/2018 0:00	Receipt and review of to and Architectural Design collaborative. Notice of Production from dated…
			Receipt and review of to and requesting copies of all documents responsive to Plaintiff's Unverified Answers to Defendant's Interrogatories.
		[illegible] 0:00	Receipt and review of to Architectural Design Collaborative from their…
			Receipt and review of to Architectural Design Collaborative from their First Request for Production to…
			Receipt and review of to Architectural Design Collaborative from their to…
			Receipt and review of to Architectural Design requesting copies from their…
			Receipt and review of to Architectural Design requesting copies from their Request for Production to…
		[illegible] 0:00	Receipt and review of pursuant to Plaintiffs' Request for Copies on…
Miss. Y	[illegible]	[illegible] 0:00	[illegible] preparation of Notice of Compliance to Request for Copies.
			[illegible], in preparation of Notice of Compliance to Request for Copies.
			Analyze defendant, response to Request for Production, in preparation of Notice of Compliance ot Request for Copies.
			[illegible] Request for Copies.
			[illegible] of Compliance to Request for Copies.
			[illegible], in preparation of Notice of Compliance to Request for Copies.
			Analyze plaintiff Answers to defendant Collateral Souce Interrogatories, in preparation of Notice of Compliance to Request for Copies.
			Analyze plaintiff Answers to defendant Loss of Consortium Interrogatories, in preparation of Notice of Compliance to Request for Copies.
			Analyze plaintiff Answers to defendant Request for Production in compliance with Medicare Secondary Payer Act, and in preparation of Notice of Compliance to Request for Copies.
			Analyze plaintiff, Answers to defendant, First Interrogatories in preparation for Notice of Compliance to Request for Copies.
			Analyze plaintiff, Supplemental Responses to defendant, First Request for Production in preparation for Notice of Compliance to Request for Copies.
			Analyze plaintiff, response to defendant, First Request for Production in preparation for Notice of Compliance to Request for Copies.
			Continue to analyze discovery pleadings in preparation of Notice of Compliance to Request for Copies.
		[illegible] 0:00	Prepare Notice of Compliance to defendants and Architectural Design Collaborative, Inc. Request for Copies dated…
			Prepare documents in compliance with Request for Copies to defendants and…
		[illegible] 0:00	Compose email to adjuster enclosing Notice of Compliance to Defendants' Request for Copies dated…
			File and serve Notice of Compliance to Defendants and Request for Copies daed…
		[illegible] 0:00	Analyze discovery responses in preparation of Notice of Compliance to Request for Copies. from…
		[illegible] 0:00	Analyze discovery responses from in preparation of Notice of Compliance to Request for Copies from regarding…
		[illegible] 0:00	Continue to review discovery pleadings in preparation of responses to Request for Copies.
Mr. X	[illegible]	[illegible]	[illegible]
			[illegible]

…then cluster items per narrative

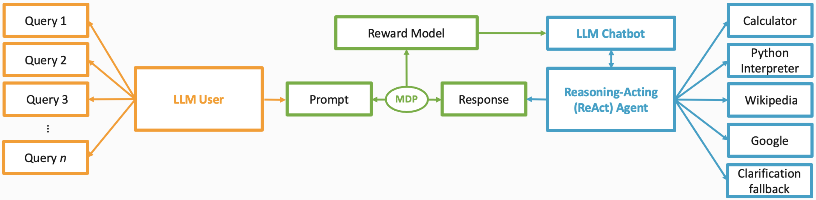
Example of 16 items at risk claimed by Mr. X:

Source	Claim ID	Date	Description	Cluster ID
Mr. X	61731	1/18/2019 0:00	Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Notice according to S. C. Code (legal document requiring signature of S. C. attorney)	1
			sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1
	61732	1/28/2019 0:00	Execute Amended Notice of Claimant's sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1
			Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Amended Notice according to S. C. Code (legal document requiring signature of S. C. attorney)	1
		10/30/2018 0:00	Advisement of medical discovery directive to (pursuant to . 67-214) non-applicability of HIPAA Privacy Rule to Workers' Compensation matters and applicable state law for evidentiary document pr oduction	1
			Advisement of medical discovery directive to (pursuant to . 67-214) non-applicability of HIPAA Privacy Rule to Workers' Compensation matters and applicable state law for evidentiary document production	1
			Advisement of medical discovery directive to EMS (pursuant to . 67-214) non-applicability of HIPAA Privacy Rule to Workers' Compensation matters and applicable state law for evidentiary document production	1
			Execute medical discovery request on through preparation of statutorily required document according to . and	3
			Initiate medical discovery on EMS through preparation of medical evidence request according to Reg. (legal document in S. C.) to obtain pertinent documents for analysis	1
			Initiate medical discovery on through preparation of medical evidence request according to Reg. (legal document in S. C.) to obtain pertinent documents for analysis	2
		10/31/2018 0:00	Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Notice according to S. C. Code (legal document requiring signature of S. C. attorney)	1
			sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1
		12/12/2018 0:00	Advisement of discovery directive to Claimant c o , . (pursuant to . 67-214) representation of employer carrier and production of evidentiary documents (L320) (A108)	1
			Initiate discovery on claimant through preparation of subpoena evidence request according to Reg. (legal document in S. C.) to obtain pertinent documents for analysis, including tax returns and business records	1
		12/5/2018 0:00	Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Notice according to S. C. Code (legal document requiring signature of S. C. attorney)	1
			sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1

Alignment of LLM Agents

Faithful AI: Learning to ask for clarifications

- **Goal**: Increase faithfulness of a given LLM in a given orchestration environment by learning when to ask for clarifications



Environment:

- Sample of *n* queries mapped to valid tools/context.
- Superalignment of LLM chatbot by another LLM:
  - **Chatbot:** Respond to prompts
  - **User:** Generate prompt and context

Reward Function:

- Every step: -1
- Ask for clarification: 0
- Tool/context requests: -5 (invalid), +5 (valid)

State:

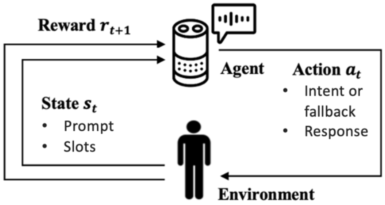
- **Chatbot**: LLM user prompt, context and tools
- **User**: Query and LLM chatbot response

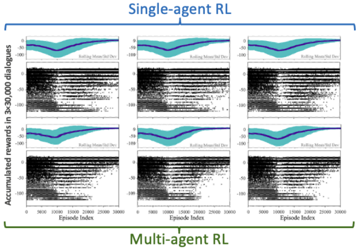
Actions:

- **Chatbot**: Select one of *k* tools and use context to respond, or fallback on asking for clarifications
- **User**: Generate semantic prompt variation for query

Faithful AI: Learning to ask for clarifications

- **Results (ICML 2023)**: Chatbot converged to policies which fulfilled intents in 99% of dialogues, in 1.8 steps on average. When users cooperated, the correct intent was fulfillied in 1.3 steps on average in 100% of dialogues





Environment:

- Library of intents with associated prompts/slots
- Single agent: Users select prompts randomly
- Two agents: User learns to select prompts.

Reward Function:

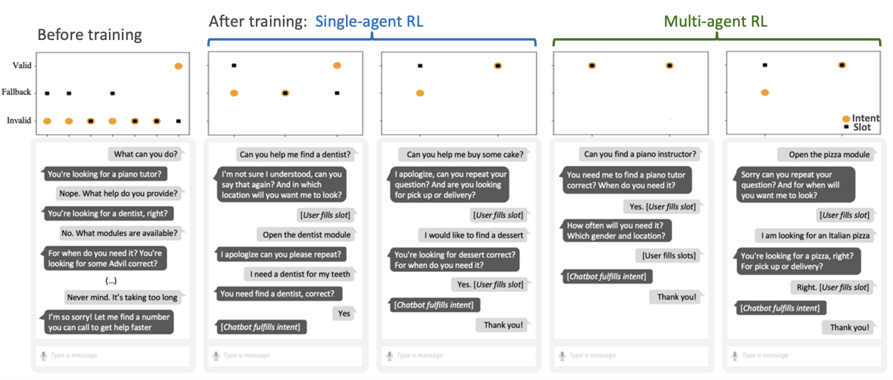
- Every step: -1
- Ask for clarification: 0
- Intent and slot guesses: -5 (invalid), +5 (valid)

Quality and efficiency of sampled dialogues

		Single RL 0-5K	Single RL 25-30K	Multi RL 0-5K	Multi RL 25-30K
% of successful dialogues		68 (.8)	99 (.1)	67 (1.4)	100 (0)
Number of steps in successful dialogues	/navigation	4.1 (.2)	1.0 (.0)	4.1 (.1)	1.0 (.0)
	/piano	4.2 (.1)	1.8 (.2)	4.1 (.1)	1.3 (.1)
	/dentist	4.2 (.0)	1.6 (.2)	4.1 (.2)	1.3 (.0)
	/pizza	4.1 (.1)	1.7 (.3)	4.3 (.0)	1.3 (.0)
	/Advil	4.3 (.2)	1.7 (.3)	4.2 (.1)	1.3 (.0)
	/dessert	4.4 (.1)	2.3 (.5)	4.3 (.1)	1.4 (.1)

The chatbot learned an original strategy...

- **Superalignment**: The chatbot found a fallback strategy to increase speed of fulfillment without sacrificing coherence: fill valid slots when prompt is ‘partially understood but too ambiguous to identify the exact intent



Your Turn!