Winter, Benjamin et al.

Conference Paper — Published Version

DDxGym: Online Transformer Policies in a Knowledge Graph Based Natural Language Environment

Suggested Citation: Winter, Benjamin et al. (2024) : DDxGym: Online Transformer Policies in a Knowledge Graph Based Natural Language Environment, In: Calzolari, Nicoletta et al. (Ed.): Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Paris, pp. 4428-4448, https://aclanthology.org/2024.lrec-main.396/

This Version is available at:
http://hdl.handle.net/11108/652

DDxGym: Online Transformer Policies in a Knowledge Graph Based Natural Language Environment

Benjamin Winter*, **Alexei Figueroa***, **Alexander Löser, Felix A. Gers, Nancy Figueroa, Ralf Krestel**
Berliner Hochschule für Technik
Christian Albrechts Universität Kiel
{Benjamin.Winter, afigueroa, aloeser, gers}@bht-berlin.de
rkr@informatik.uni-kiel.de
nfigueroa@achs.cl
*Both authors contributed equally.

Abstract

Differential diagnosis (DDx) is vital for physicians and challenging due to the existence of numerous diseases and their complex symptoms. Model training for this task is generally hindered by limited data access due to privacy concerns. To address this, we present DDxGym, a specialized OpenAI Gym environment for clinical differential diagnosis. DDxGym formulates DDx as a natural-language-based reinforcement learning (RL) problem, where agents emulate medical professionals, selecting examinations and treatments for patients with randomly sampled diseases. This RL environment utilizes data labeled from online resources, evaluated by medical professionals for accuracy. Transformers, while effective for encoding text in DDxGym, are unstable in online RL. For that reason we propose a novel training method using an auxiliary masked language modeling objective for policy optimization, resulting in model stabilization and significant performance improvement over strong baselines. Following this approach, our agent effectively navigates large action spaces and identifies universally applicable actions. All data, environment details, and implementation, including experiment reproduction code, are made publicly available.

Keywords: Reinforcement Learning, Masked Language Modelling, Knowledge Graphs, Transformers

1. Introduction

Differential diagnosis (DDx) is a crucial yet challenging task requiring doctors to sequentially narrow down potential diseases through various examinations, each revealing different symptoms. This process is complex due to the multitude of diseases with similar symptoms and the challenges associated with time, cost, and risks of different diagnostic procedures like CT-scans and laparoscopy, which might expose patients to radiation, surgery, and potential complications. Furthermore, access to powerful diagnostic tools such as an MRI might be limited, thus other avenues of examination might have to be considered first. Depending on the disease, the patient's condition might also deteriorate while different examinations are applied. Therefore, a quick diagnosis is paramount.

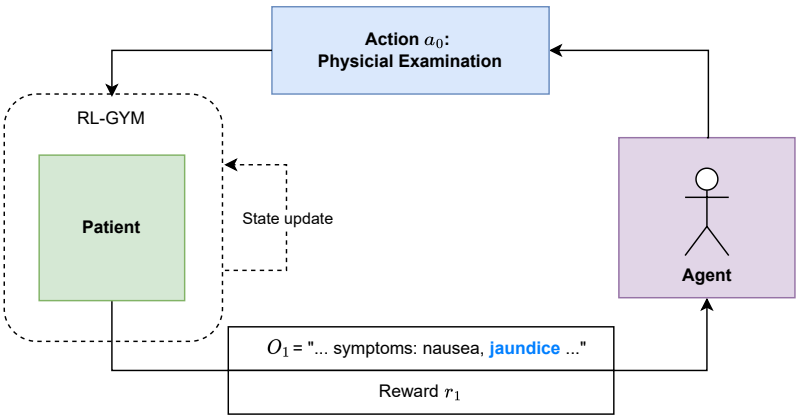
We approach the task of differential diagnosis as a single-agent reinforcement learning problem. Recent research has demonstrated the benefit of RL agents with long-term planning capabilities, both in challenging games such as chess(Silver et al., 2017) and go(Silver et al., 2016), as well as in real-world applications(Kalashnikov et al., 2021). Uniquely, we frame this task as a natural-language-based online-learning problem.

Framing differential diagnosis (DDx) this way presents an implicit set reduction problem where the agent aims to identify the patient's actual disease from possible diseases via a process of elimination. Strategic planning is essential as both disease and symptom composition are unknown, with intricate interactions between symptoms and procedures. Naively choosing examinations that eliminate the most disease candidates might not yield a diagnosis before the patient ultimately deteriorates. Therefore, considering complete trajectories given disease priors is needed.

Beyond diagnosing the patient's condition, the agent should learn to remedy *symptoms* that are particularly severe and harmful, because these might lead the patient's condition to deteriorate before the main diagnosis and treatment of the disease can be achieved. We aim to capture these dynamics in our created environment DDxGym where the agent, acting as a doctor, diagnoses and treats one such patient per episode.

Medical Text and Transformers for RL We focus specifically on creating a text-based environment for this task, since most of the information necessary for doing differential diagnosis is naturally text-based in the form of the patient's history or doctor's notes. We choose transformer language models as a baseline to solve this environment, since they have been shown to be the

LREC-COLING 2024, pages 4438–4448
20-25 May, 2024.




```
----Initial State------------------------------
List of symptoms   : nausea

----step 1 [physical examination]--------------
List of symptoms   : nausea, jaundice, fatigue
Treated symptoms   :
Applied procedures : physical examination

----step 2 [run test - bloodtest]--------------
List of symptoms   : nausea, jaundice, fatigue
Treated symptoms   :
Applied procedures : physical examination,
                     run test - bloodtest

----step 3 [use device - ct]-------------------
List of symptoms   : nausea, jaundice, fatigue,
                     liver cancer discovered
Treated symptoms   :
Applied procedures : physical examination,
                     run test - bloodtest,
                     use device - ct

----step 4-[procedure - radiotherapy]----------
List of symptoms   : ...
Treated symptoms   : liver cancer discovered
Applied procedures : ...
```

Figure 1: *Left*: Overview of DDxGym and the reinforcement learning setup. Given an initial observation that includes a symptom "nausea", the agent chooses a_0(physical examination). This results in an additional symptom "jaundice" being discovered, as seen in observation O_1 with the corresponding reward r_1. *Right*: A full example episode of our best agent interacting with DDxGym treating liver cancer.

strongest architectures for many natural language processing tasks.

However, Transformers struggle with instability in online reinforcement learning problems when acting as a policy(Parisotto et al., 2020). To address this issue, we propose a novel training approach based on an auxiliary masked language modelling objective that complements the reinforcement learning training. Using this approach, we outperform other baselines including a standard transformer, but we show that our environment remains challenging in this online setting.

Analysis with Medical Professionals We create our OpenAI-Gym based environment with data from online medical resources curated by medical professionals. We further evaluate this data and the resulting environment with medical professionals. Lastly, we perform an in-depth quantitative and qualitative analysis on trajectories of our best-performing agent to highlight strengths and shortcomings of the learned policy. In this analysis, we also show that symptom-examination overlap among diseases is a major factor contributing to the difficulty of this task.

To summarize, the contributions of this paper are as follows:

1. To the best of our knowledge, we are the first to phrase the differential diagnosis problem as an online text-based RL task. We expand this task with the treatment of the patient.

2. We create an environment and release it together with the underlying data labelled with the help of medical professionals[1].

3. We propose a masked language model (MLM) objective as concurrent loss to improve online RL with transformers: namely, environment modelling and regularisation. We show that this approach outperforms other baselines. [2]

4. We provide an in-depth qualitative analysis on the trajectories of our best model.

The remainder of this paper is structured as follows: in Section 2 we discuss related research, in Section 3 we detail both the DDxGym environment, and the knowledge graph it is based on, and in Section 4 we describe our model and training approach. Then, in Section 5 we detail our experiments, in Section 6 we discuss our experimental and analysis results, and we close with Sections 7 and 8 in examining limitations of our work, proposing future work, and our conclusions.

2. Related Work

RL in Automated Diagnosis Systems Given the interactive nature of the clinical diagnostic process, RL has been used as a suitable framework to solve it. (Tang et al., 2016) proposed an ensemble of neural networks corresponding to anatomical parts of the body which questions the patient for symptoms and diagnoses diseases. They follow up on this work by introducing hierarchical rein-

[1]The data and code for the RL environment are available at https://github.com/DATEXIS/Medical-Gym/tree/afigueroa/rayupgrade

[2]The source code to reproduce our experiments can be found at https://github.com/DATEXIS/Medical-RL/tree/afigueroa/rayupgrade

forcement learning (HRL), contextual demographics, as well as hereditary and medical history information (Kao et al., 2018). Similarly, (Yuan and Yu, 2021) decomposed the diagnostic process by aligning an RL agent trained to uncover symptoms with a classification objective for the diagnosis step. Furthermore, an automatic symptom detection system based on a graph-memory-network agent was proposed (Luo et al., 2020). In contrast to these works, we define the environment to produce only natural language observations and approach the training of the agents purely with NLP methods. Further, we add the additional process of treatment to the patient episodes and we don't make a distinction between the action spaces of examinations and treatments.

Structured Medical Knowledge Several authors have focused on creating language definitions to express the medical, structured knowledge in computer interpretable guidelines for clinical decision support systems (CDSSs) (Shiffman et al., 2001; De Clercq et al., 2001; Boxwala et al., 2004; Fox et al., 1998). We don't focus on the formalism surrounding medical knowledge representation, instead we create an RL environment. An agent trained in this environment then proposes examinations and treatments analogous to a CDSS.

Curated knowledge bases are at the core of many commercial CDSSs (Hirsch et al., 2020; Nordon et al., 2019; Razzaki et al., 2018). More generally, approaches such as UMLS (Lindberg et al., 1993) or SNOMED (Rothwell and Cote, 1996) aim to unify several biomedical concepts into abstract general-purpose knowledge graphs. Being accordingly general, these are not specific enough for the symptom-procedure relations required by DDxGym. We differ by letting an RL environment be defined by a simple, yet extensible, knowledge base that encompasses multiple diseases with very concrete edges and semantic descriptions which we make openly available.

Transformers and Reinforcement Learning Following their success in supervised settings, transformers are increasingly used in RL. Transformers used as policies face challenges such as learning stability and low sample efficiency (Li et al., 2023). (Parisotto et al., 2020) highlight the problem of stability and tackle memorization tasks with a gating architecture replacing residual connections. In our approach, we don't modify the transformer architecture but rather add a concurrent objective to stabilize the policy learning.

Generative language models can be used to initialize a policy for learning in environments adapted to yield textual observations (Li et al., 2022). Similarly, (Yao et al., 2020) use language models which are fine-tuned on human gameplay to filter for admissible actions to serve as input to a policy. These works focus on leveraging the common-sense grounding the language models acquired during pre-training, yet don't use them as policies. Our approach differs by fully modelling and learning a policy with a transformer.

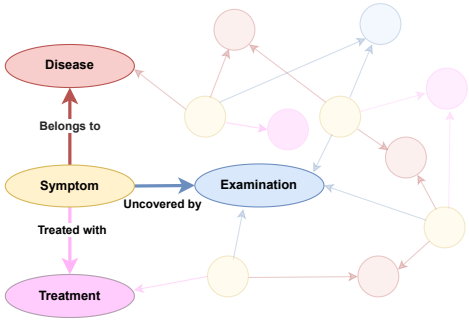


Figure 2: Disease relations in the DDxGym knowledge graph. Procedures might connect to different symptoms and therefore multiple diseases.

(Chen et al., 2021) rephrase RL as a supervised learning problem solved by an auto-regressive model of reward, states, and actions. Similarly, (Carroll et al., 2022) model reward, state, and actions with bidirectional transformer encoders and masked language modelling. In our work we don't fine-tune the LMs in a supervised setting. We train our policies exclusively in an online setting.

More recently, in systems such as InstructGPT (Ouyang et al., 2022) and ChatGPT (Schulman et al., 2022) transformers have incorporated RL mechanisms to yield impressive results in interactive settings. In contrast, we train our models directly in an online RL environment with an explicit reward, hence we don't utilize a reward approximation model(RLHF).

3. DDxGym Environment

We understand the problem of diagnosing, and subsequently curing a patient, as a partially observable Markov decision process (POMDP) described by (S, A, P_a, R_a). Figure 1 shows an overview of the DDxGym environment as well as an example episode trajectory.

3.1. Environment Definition

States and Observations Each state $s \in S$ consists of: The patient's disease $d \in D$, the patient's symptoms $z \in Z_{disease}$ with their respective states (*hidden, discovered, cured*), a main symptom z_{main}, the history of actions the agent has taken, and a decaying value H denoting the "health" of the patient.

The observation O given to the agent is a text sequence that describes the patient's *discovered* and *cured* symptoms, and the applied procedures. This emulates an electronic health record (EHR). The *hidden* symptoms, the value of H, and the disease d are not observable, which makes the environment as a whole partially observable. An example episode with such observations can be seen in Figure 1 *Right*.

Actions The action space encompasses all the procedures that are available to the agent. There is no way to distinguish actions of examination from treatment. The actual diagnosis of the disease occurs implicitly with the examinations. This is in contrast to other works (Yuan and Yu, 2021), which model the disease prediction separately. Our definition of the action space doesn't reveal any structure of the problem to the agent which makes it more challenging.

Episode Dynamics Each episode begins with a patient with one randomly sampled disease. The value of H is initialized to a positive integer, in our case 200. This is a hyperparameter that mainly determines the budget of interactions that the agent has with the patient. The initial observation includes one symptom and no procedures. This symptom is sampled by occurrence probabilities from the set of all symptoms which are not the main symptom and have an initial onset. This emulates the chief complaint i.e. the reason for the patient to visit the doctor.

Each step in the environment constitutes choosing exactly one procedure to apply to the patient. With every action, symptoms may be detected or treated, and the observations and reward are updated accordingly. With each step, regardless of which procedure the agent applies, the patient deteriorates i.e. H is reduced. How quickly the value deteriorates depends on the severity of the disease, and the severity of the untreated symptoms of the patient. This incentivizes the agent to learn a policy that treats very severe symptoms (e.g., internal bleeding), before focusing on diagnosing the underlying disease. Each episode terminates when the value of H becomes negative, or if the disease is detected and treated.

Reward Our environment features a sparse reward structure. The largest positive rewards are given only when the disease is diagnosed and subsequently treated. Smaller positive rewards are given for uncovering and treating symptoms that are not the main symptom. The value of this reward depends on whether the symptoms are diagnosed or treated, and on how severe they are. In each step where the chosen action does not lead to the detection or treatment of any new symptom, a negative reward is given with the value of the deterioration of H for this step. This steady negative reward encourages the training of policies that treat the patient in the smallest amount of steps possible, which is desirable.

Environment Concept	# distinct entities
Diseases	111
Symptoms	384
Examinations	154
Treatments	176

Table 1: DDxGym knowledge graph statistics. Examinations and treatments result in a total of 330 actions.

To summarize, the reward structure of the DDxGym environment is

$$r_t = \begin{cases} 1000 & \text{if } z_{main} \text{ is cured} \\ 100 & \text{if } z_{main} \text{ is discovered} \\ (50, 20, 10) & \text{if } z \neq z_{main} \text{ is cured} \\ (20, 10, 5) & \text{if } z \neq z_{main} \text{ is discovered} \\ \sum z_s & \text{otherwise} \end{cases} \tag{1}$$

with r_t being the step-wise reward, and $\sum z_s$ being the sum of the severity of all untreated symptoms of the patient and $z_s \in (-5, -2, -1)$. Values in parenthesis are for high, medium, and low severity symptoms respectively. This reward structure results in a lower bound of cumulative reward that is equal to $-H$ (-200), and an upper bound of 1200 for our data.

3.2. DDxGym-Knowledge Graph

In order to create the DDxGym environment as described in the previous section, we need a structured data definition that captures the medical concepts and interactions of procedures, symptoms, and diseases, i.e., a knowledge graph. Finding this type of data is challenging. While there are a few proprietary knowledge graphs available that contain this type of information, they are not freely available for research ((Nordon et al., 2019; Hirsch et al., 2020) or the work of Infermedica[3]). Widely used open medical knowledge graphs such as UMLS (Lindberg et al., 1993) and SNOMED (Rothwell and Cote, 1996) don't capture the treatment-symptom, or examination-symptom relations that would be needed to specify the environment. For that reason, we decide to label our own data and

[3] https://developer.infermedica.com/docs/v3/medical-concepts

Symptom	Examination	Treatment	Severity	Onset	Probability	Is main?
A. pancreatitis	Run test - blood lipase & amylase	IV (Fluids)	high	initial	always	yes
Fever	Physical Examination - Body temp.	antipyretics	mid	short	medium	no
Nausea	Interview - nausea	antiemetics	mid	short	medium	no
Jaundice	Interview - visual		low	short	medium	no

Table 2: *Acute Pancreatitis* in our knowledge graph. Four symptoms are associated, with the disease identifier being the main symptom. Each symptom is revealed through at least one *examination*. The *probability* field dictates which non-main symptoms appear, potentially with delayed *onset*. Though the primary treatment goal is pancreatitis, addressing *fever* and *nausea* may be crucial due to their *severity* accelerating patient deterioration. Notably, no direct *treatment* exists for *jaundice* in our data.

create a suitable knowledge graph[4].

Labelling process. Our knowledge graph is derived from educational disease resources[5] curated by medical professionals. Experts in biomedical NLP extracted and labeled *diseases*, *symptoms*, *examinations*, *treatments*, and their relations. Semantic descriptions for symptom entities were added, detailing probability of occurrence, their onset, and severity. The interaction of the entities can be seen in Figure 2. This data was refined through review and fine-tuning by medical doctors who normalized entities, disambiguating acronyms, merged duplicates and completing missing values. Table 2 displays an example of labeled entities in the graph.

Knowledge Graph Statistics Table 1 shows the results of this labelling process. After clean-up and quality assessment, the environment is comprised of 111 diseases that are diagnosed and treated with a sum total of 330 unique procedures. This represents a relatively large action space and one of the main factors that make this environment challenging. This complexity grows as more diseases and procedures are added, as is the case of highly curated commercial knowledge bases. While we are limited in this work to our knowledge graph, our approach is generally applicable to these much larger commercial alternatives.

4. Reinforcement Learning

Learning to solve DDxGym, involves solving the POMDP described in Section 3. To this end, we train an agent using online RL. Generally, DDxGym can be parallelized to yield a large amount of sampled episodes because of the low environment-step cost. This is very suitable for RL algorithms such as IMPALA (Espeholt et al., 2018) in contrast to more sample-efficient algorithms like PPO (Schulman et al., 2017). Nevertheless, our implementation of the environment and our usage of the transformer encoder are algorithm agnostic. IMPALA expects a model that encodes the observations and produces an *action distribution* and a *value estimation* of the current environment state. Since our observations are text-based, we chose a transformer language model as the encoder.

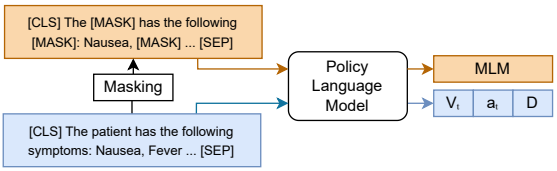


Figure 3: Model Architecture: Each environment step involves two forward passes. First, state value V_t, next action a_t, and optionally patient's disease are predicted from observation o_t. Second, the masked observation trains the MLM objective.

Transformer Encoder We encode the textual observations of the environment with a transformer. We don't require generating natural language to interact with the environment, thus we consider encoder-based language models from the BERT (Devlin et al., 2018) family. To compute the *action distribution* and *value estimation*, the [CLS] embedding of the Transformer encoder is projected through a corresponding separate linear layer.

Transformers have been shown to be unstable as RL policies(Parisotto et al., 2020; Li et al., 2023). To address this issue, rather than modifying the encoder, we add parallel learning objectives. We explore objectives where they already perform well in a (self-)supervised setting(Devlin et al., 2018): masked language modelling and text classification. The model learns these additional objectives ***concurrently*** with the reward-based IMPALA objective while interacting with the environment. Figure 3 illustrates this process.

[4]The knowledge graph is available as part of the RL environment in https://github.com/DATEXIS/Medical-Gym/tree/afigueroa/rayupgrade

[5]www.nhsinform.scot, www.mayoclinic.org, www.nhs.uk

Masked Language Modelling(MLM) This task is motivated by the aim of aligning the model to the idiosyncrasies of the medical terminology as well as the way the information is conveyed in the observations of DDxGym. With this parallel objective, we believe that we keep the model from overfitting to the purely control-based aspect of modelling DDx.

For MLM, we follow the pre-training of BERT (Devlin et al., 2018). Specifically, for each environment observation, we mask at random 15% of the tokens and task the encoder with predicting them. These predictions are then compared to the ground-truth tokens using token-wise cross-entropy, which is aggregated as the loss for this objective. Since the computation of the *action distribution* and *value estimation* require the unmasked observations, we perform two distinct forward passes, then sum the IMPALA loss and this objective's loss. This combined loss is then applied in a single backward pass.

Disease Prediction Objective We add a supervised objective for the model to classify the disease of the patient in each step. The motivation for this objective is two-fold. Firstly, we believe that it is a useful additional signal to model associations of state spaces and diseases. Secondly, this supervised objective could be used at test time for explainability: knowing which disease the agent is predicting in each step could serve as an explanation of why certain procedures are chosen. In contrast to the MLM objective, this objective does not require a second forward pass through the model, and instead we reuse the [CLS] embedding, and project it through yet another linear layer. Similar to the MLM objective, this cross entropy loss is then summed with IMPALA's loss.

5. Experiments

We evaluate the different objectives described in Section 4 using the *mean episode reward* and the *mean episode length* as metrics. Both criteria are proxies for how well the agent learned to diagnose and treat a patient. Ideally, a good agent achieves both, high mean rewards, solving a large subset of diseases, as well as low episode lengths, treating patients in an efficient manner. For our experiments we tokenize the observations and truncate sequences above 128 tokens. Due to time and resource constraints, we limit the training of our agents to a maximum of 80M steps.

Transformer Comparison We evaluate three different models to use one as our policy: BERT-base(Devlin et al., 2018), the domain adapted ClinicalBERT(Huang et al., 2019), and BERT-small (Turc et al., 2019), a compressed version of BERT. Results of these experiments are illustrated in Figure 4. All models converge to a mean reward between 100 and 200, which is poor considering that the upper bound is approximately 1200. While ClinicalBERT with its domain adaptation performs slightly better than the other two models, both ClinicalBERT and BERT-base drop sporadically to the lower bound reward of -200, highlighting their instability. BERT-small performs similarly to BERT-base, while exhibiting more stability, possibly due to the regularisation caused by the lower number of parameters. Further, with BERT-small we are able to process episodes ≈1.5x faster than with the larger models. For these reasons we chose BERT-small for subsequent experiments.

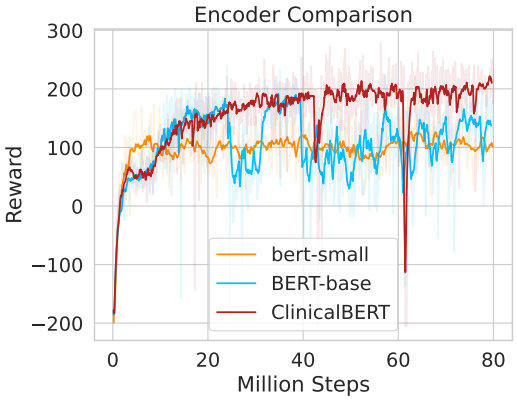


Figure 4: Different pre-trained encoders in the environment as a policy (see 5, *T*). Although all models achieve similar reward, bert-small converges with higher stability. We show EMA with $\alpha = 0.85$.

Models and Baselines We assess different baselines and model variations to evaluate the impact of our methods:

T : We employ a transformer encoder as outlined in Section 4 with only two added linear layer outputs to predict the action distribution and the value function. This basic variant excludes additional objectives, and functions as our baseline.

T+PD : We expanded *T* with one additional linear layer for the disease prediction objective.

T+MLM : We extended *T* with an additional MLM objective as described in Section 4.

T+PD+MLM : Uses both the additional supervised disease prediction and the MLM objectives.

Text-sequence LSTM : We evaluate an LSTM(Hochreiter and Schmidhuber, 1997) because of their known sequence modelling abilities and wide usage in RL. This 3-layer bidirectional LSTM, with dropout and 256 hidden layer size, mirrors the vocabulary and word embedding layer of the transformers for fair comparison.

Random : We evaluate this fully random policy to set a lower bound.

6. Results and Discussion

6.1. Quantitative Results

Figure 5 shows the results of evaluating the experiments previously described. All of our baselines beat the random policy by a wide margin in mean episode reward. However, some of the compared models, *T* in particular, barely achieve shorter episode lengths. That means these models learn to diagnose and treat symptoms, but fail to actually cure the disease. *T+MLM* significantly outperforms all models in both reward and episode length. We also observe this model's reward is significantly more stable than *T+PD*, *T+MLM+PD*, and the *LSTM* models. Additionally, the mean episode length is the most stable in comparison to the other models. While the *LSTM* outperforms BERT-small without any auxiliary objectives, it falls short of *T+MLM*. This is an interesting result since transformers have superseded LSTMs in supervised learning, the RL setting keeps being challenging and DDxGym is no exception. The additional disease prediction objective in *T+PD* and *T+MLM+PD* leads to increased instability, and did not achieve improved performance. We believe that this objective distracts the agent from the actual control problem of diagnosing and treating.

6.2. Qualitative Results

We simulate 5000 episodes of inference with a checkpoint of T+MLM that achieves the highest mean reward. To group the diseases we use episode lengths as they can be easily discretized and, as shown by previous results, strongly correlate with the reward.

Successful diseases We examine the distribution of the diseases with respect to the episode lengths Figure 6 *Left*. We notice that the agent achieves the near maximal reward for episodes with 6 steps or less. Threre are 50 such diseases, almost half of the total. We expand on the action distributions of this group of diseases at the top row of Figure 6 (Right). The agent learns to uniquely use examinations in the first step (only blue actions), while in the intermediate steps the agent continues trying diagnostic actions to finally treat the main symptom in the last step of the episode (only magenta actions). Thus, for these 50 diseases the agent learns the correct trajectories of examining symptoms, considering diagnostics, and proposing treatments. We show one such successful episode of our agent treating liver cancer in Figure 1.

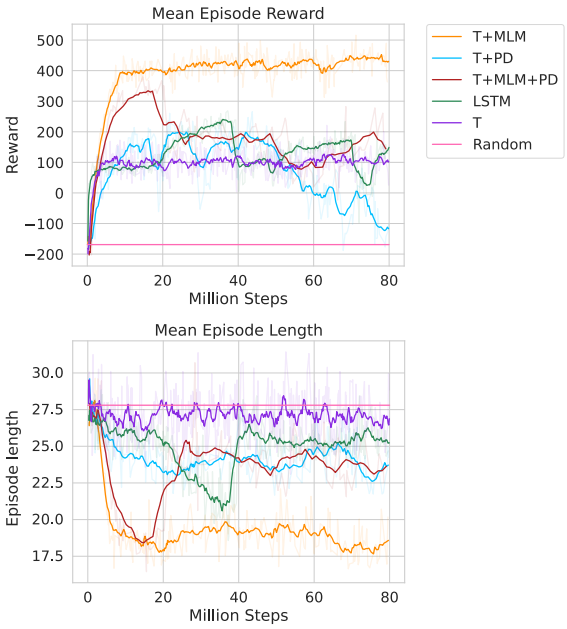


Figure 5: Five policies on DDxGym, (T+MLM) outperforms in mean reward (top), episode length (both), and learning stability treating patients quicker and more successfully.

Notably, the agent *learns to discriminate* between diagnostic and treatment actions even though they are not explicitly distinguishable within the action space. It is also noteworthy that the agent learns to treat the disease only once the main symptom is uncovered, which is evident since the intermediate steps involve mostly diagnostic actions. We believe that this *awareness* regarding the action space and episode state is enabled by the coexisting MLM objective while training, since via this mechanism recall of past episodes is tightly coupled to the reward.

Unsuccessful diseases For episodes longer than 6 steps the rewards are mostly negative (second row of Figure 6 Right), which means the agent is not succesful at uncovering nor treating the disease. We highlight how for these diseases there's no discrimination by the agent of diagnosis and treatment actions in any of the steps (both magenta and blue actions are present).

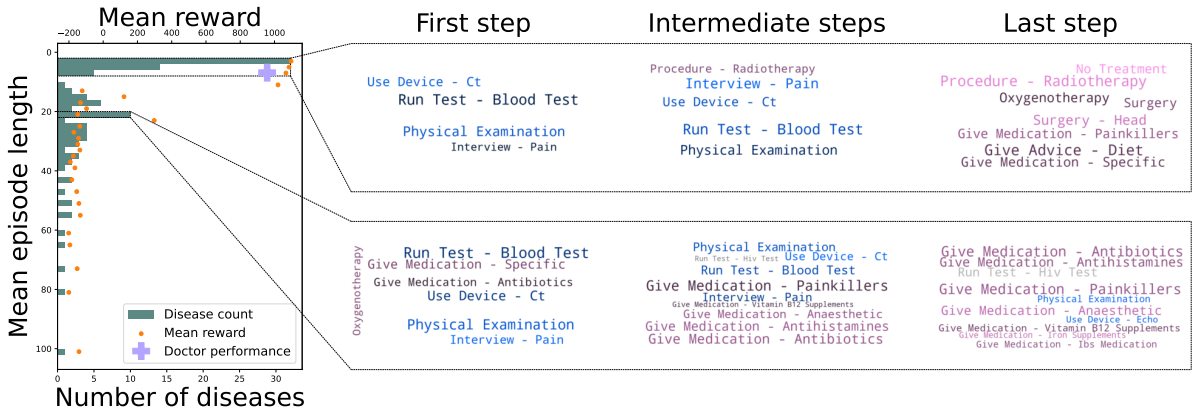


Figure 6: Inference on 5000 episodes with the best *T+MLM model*. *Left*: distribution of diseases across episode lengths. For 50 diseases the agent solves the environment in under 6 steps. A high mean reward (orange) correlates to a short episode length. Doctor performance shown for comparisson. *Right*: Action distributions of episodes solved under 6 steps (top), and in $[19, 20]$ steps (bottom). For solved diseases the agent learns to uncover symptoms initially (blue) and then follows these with treatments(magenta).

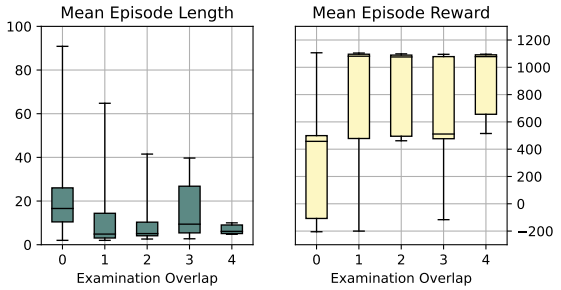


Figure 7: Mean episode lengths (left) and rewards (right) distribution based on overlap in examination actions revealing the main symptom. Diseases with higher overlap have shorter episodes and higher rewards, while those with low overlap present longer episodes, indicating that diseases with narrowly applicable diagnostic actions are more challenging to diagnose and treat.

Examination Overlap To investigate what makes certain diseases successful or unsuccessful we analyzed the examinations that uncover their main symptoms. We construct a disease-pairwise comparison for the simulated episodes. We group the diseases with respect to the size of the intersection of their examinations and denote this the *examination overlap*. Figure 7 *Left* shows the distributions of the mean episode lengths for these overlap groups. In the same manner we examine the behavior of the mean reward in Figure 7 *Right*. We note that the groups with higher overlap present both lower episode lengths as well as higher rewards. The decrease of the examination overlap leads to a higher episode length, a lower reward, and a noticeably higher variance for both. We believe that this relation of the episode length-/reward and main-examination overlap supports the idea that diseases are particularly challenging, when the examinations required to diagnose them are incresingly specific. We have similar findings evaluating the environment with medical doctors.

Human Expert Trajectories We sample the performance of a medical doctor in DDxGym for 16 diseases, marking the mean reward and mean episode length in Figure 6. This experiment shows a medical professional can achieve almost ideal rewards even without prior experience with the environment. The doctor identified and used many of the same actions that were also discovered by our best agent as being the most generally applicable, such as running a generic blood test, and physical examinations. However, as episodes went longer, the doctor was much more capable than our agent of choosing procedures that complement each other to further narrow the set of possible diseases, failing only in one episode.

7. Limitations and Future Work

Despite aiming to realistically model differential diagnosis, resource constraints necessitated simplifications. These limitations, mostly relate to data collection and granularity, and can be addressed for enhancement of the DDxGym environment. For example, the current version only incorporates text modality, ignoring complex, multimodal patient data like lab results and images, and it does not account for disease interactions or comorbidities. This would require further data labelling and models beyond transformer encoders.
Additionally, incorporating risks and costs associated with examinations and treatments is essential. This would also necessitate additional labels and

reward structure adjustments, but it would lead the learning policy to more strongly consider patient comfort, risks, and costs, opting for specific actions only when imperative. Finally, improving the observation generation process through enhanced templating or generative models is crucial due to the challenging nature of Electronic Health Records (EHRs), which in the natural language sense are complex and heterogeneous.

8. Conclusion

We present DDxGym, a novel text-based reinforcement learning environment for the core medical task of differential diagnosis. In addition, we provide for the community a medical knowledge graph with 111 diseases and their symptoms, procedures, and their interactions. Further, we propose a novel masked language modelling objective to address problems of learning instability and the overall performance of transformer language models in online RL. This approach significantly outperforms a regular pre-trained transformer and other baselines. Our analysis shows that our agent approaches Differential Diagnosis in a similar way as our medical expert discovering generally effective examinations, discriminating among procedures and remedying secondary symptoms.

9. Ethics statement

Differential diagnosis is a challenging and complex problem, with a direct and immense impact on human lives. While research in this direction may yield promising results in terms of accuracy, this is not sufficient for these models to be applied in the real world. The major ethical and moral hurdles that will have to be overcome are transparency, accountability, and fairness. Much more research has to be done in making sure that such models can be interpreted, and that they are unbiased, to understand their exact limitations and where and why they fail. We also believe, that this research should not be done to replace medical professionals, and only to support them.

10. Aknowledgements

We would like to thank the editors and anonymous reviewers for their helpful suggestions and comments. This work was funded by by the German Federal Ministry of Education and Research (**BMBF**) under grant agreements 01IS23013C (**More-with-Less**), as well as the grant agreement 01IS23015A (**AI4SCM**) and the grant agreement 16SV8857 (**KIP-SDM**) as well as the **Medical AI Analytics & Information GmbH**.

Aziz A Boxwala, Mor Peleg, Samson Tu, Omolola Ogunyemi, Qing T Zeng, Dongwen Wang, Vimla L Patel, Robert A Greenes, and Edward H Shortliffe. 2004. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *Journal of biomedical informatics* 37, 3 (2004), 147–161.

Micah Carroll, Orr Paradise, Jessy Lin, Raluca Georgescu, Mingfei Sun, David Bignell, Stephanie Milani, Katja Hofmann, Matthew Hausknecht, Anca Dragan, et al. 2022. Unimask: Unified inference in sequential decision problems. *arXiv preprint arXiv:2211.10869* (2022).

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

Paul A De Clercq, Arie Hasman, Johannes A Blom, and Hendrikus HM Korsten. 2001. Design and implementation of a framework to support the development of clinical guidelines. *International journal of medical informatics* 64, 2-3 (2001), 285–318.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*. PMLR, 1407–1416.

John Fox, Nicky Johns, and Ali Rahmanzadeh. 1998. Disseminating medical knowledge: the PROforma approach. *Artificial intelligence in medicine* 14, 1-2 (1998), 157–182.

Martin Christian Hirsch, Simon Ronicke, Martin Krusche, and Annette Doris Wagner. 2020. Rare diseases 2030: how augmented AI will support diagnosis and treatment of rare diseases in the future. *Annals of the Rheumatic Diseases* 79, 6 (2020), 740–743. https://doi.org/10.1136/annrheumdis-2020-217125 arXiv:https://ard.bmj.com/content/79/6/740.full.pdf

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *CoRR* abs/1904.05342 (2019). arXiv:1904.05342 http://arxiv.org/abs/1904.05342

Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. 2021. MT-Opt: Continuous Multi-Task Robotic Reinforcement Learning at Scale. *CoRR* abs/2104.08212 (2021). arXiv:2104.08212 https://arxiv.org/abs/2104.08212

Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, Jacob Andreas, Igor Mordatch, Antonio Torralba, and Yuke Zhu. 2022. Pre-Trained Language Models for Interactive Decision-Making. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 31199–31212.

Wenzhe Li, Hao Luo, Zichuan Lin, Chongjie Zhang, Zongqing Lu, and Deheng Ye. 2023. A Survey on Transformers in Reinforcement Learning. *arXiv preprint arXiv:2301.03044* (2023).

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics* 2, 01 (1993), 41–51.

Hongyin Luo, Shang-Wen Li, and James Glass. 2020. Knowledge Grounded Conversational Symptom Detection with Graph Memory Networks. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 136–145.

Galia Nordon, Gideon Koren, Varda Shalev, Eric Horvitz, and Kira Radinsky. 2019. Separating Wheat from Chaff: Joining Biomedical Knowledge and Patient Data for Repurposing Medications. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 9565–9572. https://doi.org/10.1609/aaai.v33i01.33019565

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. 2020. Stabilizing transformers for reinforcement learning. In *International conference on machine learning*. PMLR, 7487–7498.

Salman Razzaki, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Michael Taliercio, Mobasher Butt, Azeem Majeed, et al. 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv preprint arXiv:1806.10698* (2018).

David J Rothwell and RA Cote. 1996. Managing information with SNOMED: understanding the model.. In *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 80.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347

J Schulman, B Zoph, C Kim, J Hilton, J Menick, J Weng, JFC Uribe, L Fedus, L Metz, M Pokorny, et al. 2022. ChatGPT: Optimizing language models for dialogue.

Richard N Shiffman, Abha Agrawal, Aniruddha M Deshpande, and Peter Gershkovich. 2001. An approach to guideline implementation with GEM. In *MEDINFO 2001*. IOS Press, 271–275.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, L. Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529 (2016), 484–489.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *CoRR* abs/1712.01815 (2017). arXiv:1712.01815 http://arxiv.org/abs/1712.01815

Kai-Fu Tang, Hao-Cheng Kao, Chun-Nan Chou, and Edward Y Chang. 2016. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS workshop on deep reinforcement learning*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation. *CoRR* abs/1908.08962 (2019). arXiv:1908.08962 http://arxiv.org/abs/1908.08962

Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. 2020. Keep CALM and Explore: Language Models for Action Generation in Text-based Games. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8736–8754.

Hongyi Yuan and Sheng Yu. 2021. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *arXiv preprint arXiv:2112.00733* (2021).