

Navigating WebAI: Training Agents to Complete Web Tasks with Large Language Models and Reinforcement Learning*

Extended Abstract[†]

Lucas-Andrei Thil
Maastricht University
the Netherlands

Mirela Popa
Maastricht University
the Netherlands

Gerasimos Spanakis
Maastricht University
the Netherlands

l.thil@student.maastrichtuniversity.nl mirela.popa@maastrichtuniversity.nl jerry.spanakis@maastrichtuniversity.nl

ABSTRACT

Recent advancements in language models have demonstrated remarkable improvements in various natural language processing (NLP) tasks such as web navigation. Supervised learning (SL) approaches have achieved impressive performance while utilizing significantly less training data compared to previous methods. However, these SL-based models fall short when compared to reinforcement learning (RL) approaches, which have shown superior results. In this paper, we propose a novel approach that combines SL and RL techniques over the MiniWoB benchmark to leverage the strengths of both methods. We also address a critical limitation in previous models' understanding of HTML content, revealing a tendency to memorize target elements rather than comprehend the underlying structure. To rectify this, we propose methods to enhance true understanding and present a new baseline of results. Our experiments demonstrate that our approach outperforms previous SL methods on certain tasks using less data and narrows the performance gap with RL models, achieving 43.58% average accuracy in SL and 36.69% when combined with a multimodal RL approach. This study sets a new direction for future web navigation and offers insights into the limitations and potential of language modeling for computer tasks.

CCS CONCEPTS

• **Computing methodologies** → **Machine translation; Planning with abstraction and generalization; Deep belief networks**; • **Human-centered computing** → **Natural language interfaces; Graphical user interfaces**.

KEYWORDS

Large Language Models, Machine Learning, Web, User Interfaces

ACM Reference Format:

Lucas-Andrei Thil, Mirela Popa, and Gerasimos Spanakis. 2024. Navigating WebAI: Training Agents to Complete Web Tasks with Large Language Models and Reinforcement Learning: Extended Abstract. In *Proceedings of*

ACM SAC Conference (SAC'24). (SAC'24), Avila, Spain, Article 4, 9 pages. <https://doi.org/10.1145/3605098.3635903>

1 INTRODUCTION

Neural networks have been used to complete web navigation tasks using multimodal inputs such as Humphrey et al [15] which combined visual and text inputs from a web page. Others focused on pure text and expanded over instruction mapping methods [13, 20, 24] but were very constrained in their language understanding tasks. The application of language models showed greater capabilities in completing computer tasks, but were all largely constrained to their environment and had poor transfer abilities. Lately, the Cambrian explosion of large language models opened the way for more capable models requiring less data and offering better transfer capabilities as shown by Gur et al. and Kim et al. [12, 16]. Even though they achieved outstanding results over different web navigation benchmarks such as Miniwob++, we highlight in this work that they are in fact more limited than previously claimed. We propose several approaches to rectify some of their issues with memorizing information and highlight the challenges of combining LLMs with a joint multimodal representation of their environment.

Our research primarily focuses on optimizing small-scale models such as grounding them more thoroughly to their environments, in order to benchmark their capabilities before contemplating further scaling. This prioritization aligns with the imperative for fast inference times and adaptability to limited or novel data scenarios.

In this paper, we evaluate the limitations of the Miniwob++ benchmark and the recorded episodes used in previous works in order to present different processing paths to correct their shortcomings. We reproduce the previous results obtained in the literature focusing in the applications of LLMs for web navigation and show that they overfit their benchmark and are unable to recover in slight changes in their environments, hampering the claims that they could understand HTML content. We test our techniques on newly trained models that we enhance by devising hierarchical planning abilities that showcase superior results. Then, we combine them with a multimodal representation using visual inputs that we trained in a supervised manner, and in a reinforcement learning one over multiple phases. We present the abilities of these models to learn such representations, but they suffer from transfer abilities due to the nature of the architectures used.

This paper makes several key contributions. First, we provide insights into the capabilities of agents trained in user interactions to adeptly navigate diverse web interfaces. Second, we introduce a more robust evaluation that exposes the shortcomings of these

^{*}Produces the permission block, and copyright information

[†]The full version of the author's guide is available as `acmart.pdf` document

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SAC'24, April 8 – April 12, 2024, Avila, Spain

© 2024 Association for Computing Machinery.

ACM ISBN 979-8-4007-0243-3/24/04...\$15.00

<https://doi.org/10.1145/3605098.3635903>

models due to their tendencies to memorize a lot from their environment and how to overcome them. We also set more accurate and grounded results over the Miniwob++ benchmark by taking into account the tendencies of previous work to overfit by applying attacks to the environment. As we showed that previous works tend to learn the distribution of target elements, we provide directions on how to correct it. Lastly, combined with the improvements over our models we deliver a comprehensive analysis of current methods' limitations in order to explore more performant architectures, further contributing to the field toward more capable approaches.

2 RELATED WORK

Recent advancements in AI-driven web navigation have led to various approaches to enhance intelligent agents' performance and capabilities. This section highlights key methods and techniques that have shaped this project.

2.1 OpenAI Universe

OpenAI's Universe, released in 2016, aimed to develop a general-purpose agent for tasks like video games and web navigation, focusing on Miniwob, a framework with over 80 embedded tasks [22][29]. Miniwob serves as a reinforcement learning environment, allowing controlled experimentation with web navigation complexities.

2.2 Neural Network Oriented Approaches

Several neural-network-oriented models have been devised for web navigation. Notably, Humphrey et al.'s CC-Net combines RL with supervised learning, achieving state-of-the-art results on the Miniwob Benchmark but requiring 2.4 million examples [15]. Liu et al.'s Workflow Guided Exploration, which constrains actions during RL training, has also shown promise [20].

Older works, such as Pasupat et al. [24], focused on traditional neural network approaches, while He et al. [13] outlined limitations in multimodal environments. These works have identified challenges and achieved successes but still face generalization issues.

The tradeoff between exploration and exploitation in RL is a common challenge, often requiring extensive data to converge to optimal solutions. While some models, like CC-Net, approach human capabilities, the need for large amounts of data highlights the importance of investigating more efficient models.

Figure 1 illustrates the average accuracy of different models over the Miniwob benchmark, comparing training techniques such as SL, RL, combined approaches, and few-shots prompting examples.

2.3 Large Language Models

Early works in web navigation with large language models (LLMs) include Nakano's 'WebGPT' [21], which used GPT-3 [8] as a browsing assistant but was limited by not using raw web content. Yao et al.'s 'WebShop' [32] created an e-commerce environment for language models but faced restrictions in context window and adaptability.

Attempts to pre-train LLMs on HTML content struggled with benchmarking on classical NLP tasks, limiting their use in web navigation [2]. However, breakthroughs like Gur et al.'s work on "Understanding HTML with Large Language Models" [12] demonstrated the potential of LLMs in web content comprehension, surpassing

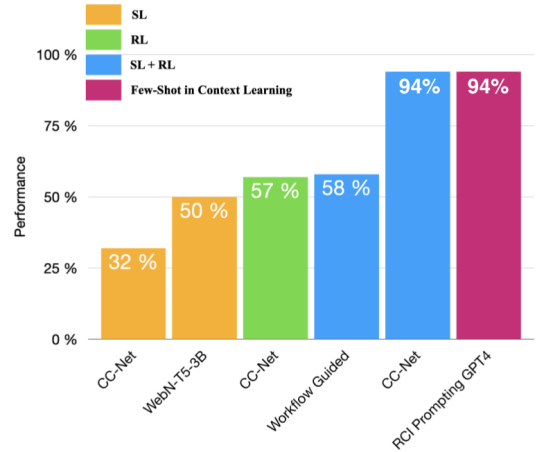


Figure 1: Comparison of Existing Models Regarding the Web Navigation Task over the Miniwob Benchmark. Average comparison over the tasks proposed on the Miniwob benchmark between different training techniques and architectures [12, 15, 16, 20].

previous supervised learning (SL) approaches with significantly less data.

Kim et al. further showcased the potential of larger LLMs like GPT4 [16, 23] in web navigation tasks through iterative prompting, achieving state-of-the-art baselines in a few-shot manner.

Despite these advancements, challenges remain in developing efficient and adaptable models for web navigation. Existing methods often require extensive training data and may struggle with fast inference times, highlighting the need for smaller, more capable models.

3 METHODS

Our methodology addresses the limitations in previous large language models (LLMs) for web navigation using the Miniwob benchmark. We begin by describing the environment and dataset, then move to model design in two stages.

In the first stage, we analyze various T5-based models, fine-tuning them with hierarchical planning techniques to overcome identified limitations. In the second stage, we integrate the best-fine-tuned model with a multimodal neural network, using both supervised learning (SL) and reinforcement learning (RL) to enhance performance and adaptability.

We also conduct an ablation study to understand the models' inner workings and identify areas for improvement. Our approach emphasizes assessing performance at a small scale before integrating additional techniques.

3.1 Miniwob++ Benchmark and Datasets

The Miniwob++ benchmark [20] offers over a hundred web-based environments, simulating web exploration scenarios. We utilize thirteen thousand human-made demonstrations provided by the Farama Foundation, enabling supervised training. The benchmark's

alignment with existing research and compatibility with reinforcement learning (RL) through the gymnasium (previously GYM) environment [7] and techniques makes it suitable for our study.

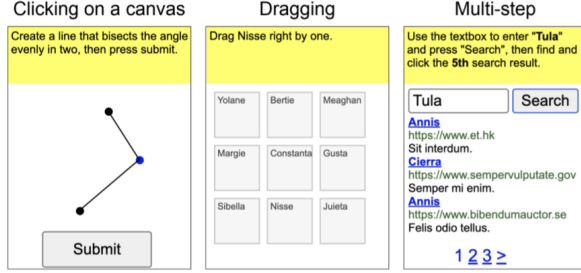


Figure 2: Example of Miniwob Episodes. Each opened episode is timed and alongside it, a discounted reward is computed. These episodes cover a wide range of tasks, and in our case, we select a subset of 40 episodes that are suited to work with language models in the fashion of Gur et al. [12].

Our datasets include HTML and Document Object Models (DOM) elements parsed as dictionaries, with unique reference numbers ('ref') for identification. We also process mouse interaction data into two actions: 'click' or 'type_text'. Unlike previous works[15], we concatenate adjacent typing actions into single actions, see Figures 3 and 4. This approach aligns with Gur et al.[12], using only two separate actions for efficiency and ease of implementation.

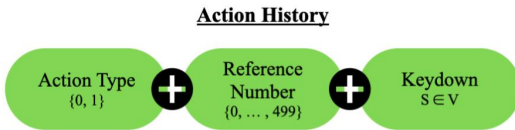


Figure 3: Structure of Action History as Proposed by Gur et al. [12].

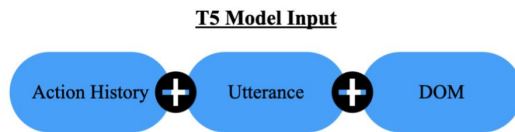


Figure 4: Structure of T5 Input in its traditional form as Proposed by Gur et al. [12].

Our methodology processes and fine-tunes data from the Miniwob benchmark for web navigation tasks. We create a hierarchical T5 model by identifying sub-tasks within episodes and translating high-level instructions into actionable plans [6]. The model is then fine-tuned for planning and action tasks [31], as shown in Figure 5.

- (1) Infer Model Plan: Devise a plan for the following instruction: 'Click the menu button, and then find and click on the item labeled next'
Subtasks = ['Click menu button', 'find an item labeled next', 'click on the next icon']
- (2) Loop through the subtasks utterances and infer model: {Action History, subtask utterance, DOM} → {Action output, reference, keydown string, boolean final state}
- (3) Perform action of the Miniwob environment: Perform action, check if the episode is terminated, and reward

Figure 5: T5 Hierarchical Inference Process. We first infer the model to devise a navigation plan from the initial utterance, then iterate through the subtask instructions individually to infer the current action at each time step while evaluating the state of the episode by means of the computed reward and terminal state.

3.1.1 Episode Processing for Action History Extraction. We process episodes to extract action history, following key steps as seen in Figure 5. The main steps consist of removing duplicate actions, and retaining only the last occurrence. Unnecessary actions, such as clicks on the `< body >` element, are discarded. Only the last keydown action for each targeted element is retained, with specific cases considered for various interactions. Manual adjustments are made if needed. To address task distribution imbalance, we down-sample 150 episodes in over-represented task suites, as seen in Figure 15.

3.1.2 Task Planning Dataset. We identify and transform various types of sub-actions, such as clicking or selection-based actions. A general function targets each Miniwob episode to produce a list of sub-tasks based on observed interactions. Specific challenges in tasks like flight booking or email forwarding are addressed, with some episodes dropped to ensure clarity. The complexity of some tasks suggests that larger pre-trained language models may perform better.

Here are two examples of the translation between Miniwob utterances and an action sequence for the hierarchical planning task:

- Example 1:
 - Utterance: "Departure City":"Philadelphia","Destination City":"Charlotte","Ticket Type":"Return flight","Departure Day":4,"Returning Day":26,"Passengers":2
 - Action sequence: Select Departure City Philadelphia; Select Destination City Charlotte; Select the Departure Day to 4;
- Example 2:
 - Utterance: Expand the section below and click submit.
 - Action sequence: Expand the section below; click submit;

The final dataset is composed of over eight thousand action episodes, and in Figure 15 we describe the number of examples per task. The second dataset regarding task planning instruction contains 10,960 episodes.

3.2 Models

In this section, we delve into an array of models designed to tackle web navigation tasks, featuring the WebN-T5 model as our cornerstone, which is based on the T5 architecture and has demonstrated superior performance due to its bi-directional attention encoder [12, 25]. Alongside WebN-T5, we explore variations such as T5-Hierarchy (T5H) fine-tuned for hierarchical tasks, and our hybrid model, CC-Net5, which combines elements from both T5 and CC-Net by Humphrey et al. [15]. These models have been trained using different strategies like supervised learning (SL) and reinforcement learning (RL), offering a rich comparative landscape for evaluating their effectiveness and limitations in web navigation.

3.2.1 WebN-T5. We attempt to reproduce the models from 'Understanding Large Language Models' [12] but face challenges with reference numbers and element distribution. As seen in Figure 6, the non-uniform distribution of elements may lead to bias or memorization [9] based on their location, or even worse entirely over the distribution of the salient elements of the page [3]. We conduct an experiment with ordered and randomized references to investigate this issue. To mitigate this, we randomize the reference numbers in all episodes, forcing the models to base predictions on element features.

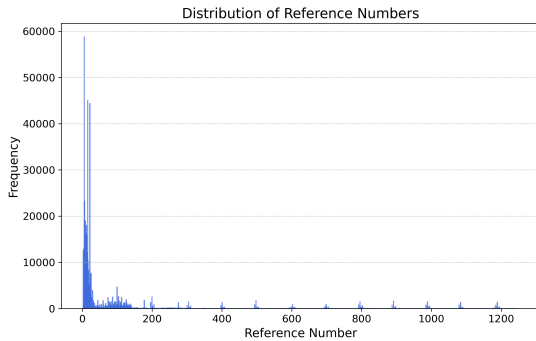


Figure 6: Distribution of Ordered Reference Numbers in the Recorded Actions. We can observe that the distribution of the target elements is concentrated among several locations, which can be linked to the salient elements in the DOM and displayed on the page. One of our claims is that previous works focused on learning these distributions by overfitting.

The T5 model was fine-tuned using ROUGE loss metrics [19], which measure the overlap between predicted and reference sequences. However, there are limitations:

- ROUGE metrics are primarily designed for text summarization or translation, not action sequence prediction.
- They do not account for the temporal order and structural dependencies in action sequences [26].

Thus, while ROUGE offers insights into the quality of generated action sequences, it may not fully capture their correctness and effectiveness.

3.2.2 WebN-T5 Hierarchical Planning. We aim to train the agent on lower-level tasks during an episode, addressing observed non-optimal actions in the original WebN-T5 [12]. The model is fine-tuned to divide tasks hierarchically, proposing a multi-step navigation plan by being trained on both datasets containing the task planning and action episodes using a supervised learning approach.

3.2.3 Multimodal Language Model with Reinforcement Learning. The best multimodal model used over the Miniwob++ benchmark is CC-Net by Humphrey et al. [15] which achieved human-level performance by using reinforcement learning. Their model used a PPO-based algorithm [28], V-MPO [30], for reinforcement learning (RL) which aims to improve performance without requiring extensive exploration. We used a similar learning approach, except that our model architecture combines a CC-Net-inspired model with a fine-tuned T5 model for the hierarchical planning task detailed earlier. The motivation behind using reinforcement learning, is that many of the recorded examples are not sufficient in covering the variety of the cases proposed. Some environments require further training, and as the exploration space is very large, a method such as V-MPO is better suited to that perspective.

The architecture is derived from CC-Net, using a multimodal approach that includes predictions from the fine-tuned T5 model, a screenshot of the current environment, and language information. The architecture details can be seen in Figure 7 and are as follows:

- Screenshot inputs are processed through four RESNET blocks.
- Language inputs are embedded and passed through a transformer layer.
- Outputs are fed into a multimodal block, concatenated with previous actions, and processed through an LSTM block.
- The final layer includes binary variables for action types and tensors for reference numbers and vocabulary indexes.

We adapt the architecture to deal with the vanishing/exploding gradient problem and use one-hot encoding for the experiment.

The design of the CC-Net5 architecture's loss function was an iterative process, influenced by the nature of its output layer. The final output layer consists of:

- Action: A binary value represented as a tensor of size 1.
- Reference number: A tensor of length 500.
- Keydown text: A tensor of size 8x1591 (8 times our vocabulary size).

Initially, a mean-squared-entropy (MSE) loss function was used, but due to sparse encoding and convergence issues, it was updated to a cross-entropy (CE) loss. This change involved predicting reference numbers and keydown text through a softmax function over each section of the output layer. The final loss function efficiently processes this output, with the action type tensor matched with its corresponding boolean value, and the reference number and keydown text token indexes retrieved from the tokenizer.

When using V-MPO, we sample actions from a normalized categorical distribution based on the activation weights of the final layer, where each index represents the token position in our vocabulary. This method allows us to derive a probability distribution for sampling actions during policy inference for exploration. The architecture involves a two-stage process, training the CC-Net-based architecture with V-MPO, while the T5 model remains static.

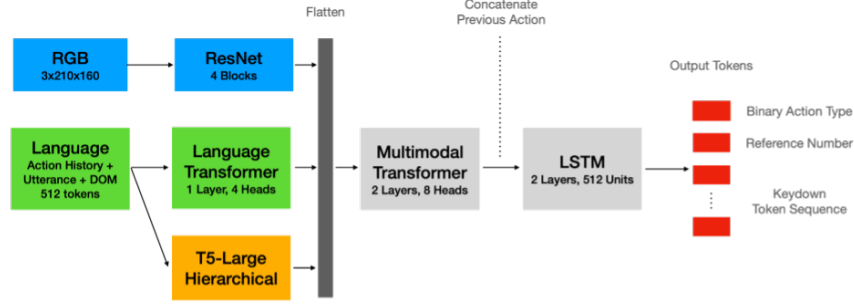


Figure 7: The Combined Architecture of T5-large fine-tuned over the Hierarchical Task, and CC-Net Multimodal Abilities over an RL Approach.

Although the CC-Net part serves as an RL boost to the original T5 model, this approach may have limitations if the original T5 inference is severely flawed.

The model’s accuracy is measured in an online environment over the Miniwob benchmark. During training, we alternate between offline (SL) and online (RL) phases [1], using recorded Miniwob episodes for the first offline phase and successful episodes from online RL phases for subsequent offline ones [27].

We propose a new preset in Gymnasium’s RL environment, reducing the action space to two types and adjusting the observation space. The time limit for episodes is increased to thirty seconds, and a discounted reward is computed to train the models in an RL manner.

4 RESULTS

This section analyzes the outcomes of various trained models, including ablation studies, and compares them with existing models. The models are benchmarked on the Miniwob benchmark, focusing on click and typing actions, and the accuracy is measured over a hundred episodes.

4.1 Model Performances

Our experiments reveal that the T5-large model, fine-tuned on hierarchical tasks, achieves superior performance with an average accuracy of 43.58%. This outperforms the T5-base model, which reaches an average accuracy of 39.78%. Interestingly, a hybrid approach combining T5-large with a CC-Net-inspired architecture yields an accuracy of 36.39% in its BC phase. However, this drops to 33.86% after the RL phase, a phenomenon we discuss in subsection 4.4.2 due to a covariate shift towards the T5 model. A comparison of the different models’ performance metrics are presented in Figure 13, including the ablation study over their inputs.

4.1.1 T5-Model Performance. Fine-tuned T5 models show variable performance. T5-large scores 43.49% accuracy on hierarchical tasks, while its ablated version slightly edges it with 43.58%. In contrast, T5-base lags with 39.77% accuracy as seen in Figures 13a and 13b. The performance advantage in tasks like ‘click-checkboxes-soft’ indicates that larger models capitalize better on their pre-trained linguistic skills.

Our replication of WebN-T5 by Gur et al. [12] found that model performance hinges on reference ordering. Training with randomized references improved its performance, validating our randomization process, as depicted in Figure 11.

4.1.2 Evaluation of Original Papers. Our findings upon reproducing Gur et al.’s work [12] reveal memorization tendencies rather than genuine task understanding. Randomizing references resulted in performance drops, questioning the original claims but reaffirming the importance of data randomization in model training.

4.2 History versus No History

An ablation study on action history reveals interesting insights. The T5-base model’s performance drops by nearly two percentage points when action history is removed. In contrast, models fine-tuned on hierarchical tasks, such as T5-large, show almost no performance change, as detailed in Table 1 and Figures 13a and 13b. This suggests that the hierarchical nature of the task allows the model to plan its actions more effectively, making it less reliant on action history. Furthermore, the randomization of references seems to make the model less dependent on previous sequences, providing another layer of resilience to the removal of action history.

4.3 Hierarchical Planning Improvements

Fine-tuning a model over the hierarchical task achieved higher results than the compared WebN-based models. The following T5 sizes have been fine-tuned which are T5-base and T5-large, achieving respectively 42.1% and 43.49% accuracy shown in Table 1. The original WebN achieved 46.4% accuracy over WebT5-large, and the reproduced T5-base size achieved 38.1% which shows that hierarchical planning is an important component of these models.

The ablation study over the action history outlined a 14% decrease in performance over the T5-base model for hierarchical planning and 0.2% over the T5-large. This is much less than the reported 6.4% reported by the original WebN-T5-large and WebN-T5-3B models showing that hierarchical planning is less sensitive to the action history as it tries to solve the episode by following an original plan step by step.

Nonetheless, we do observe significant rates of failures in complicated episodes, or when the environment changes as the agent follows the original plan as seen in detail in Figures 13a and 13b

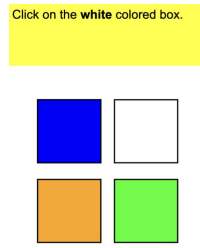


Figure 8: Episode example of choosing one out of four colors displayed from the original utterance.



Figure 9: Episode example of checking boxes of words related to the one given in the original utterance.

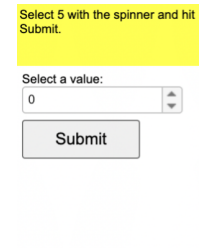


Figure 10: Increment the spinner to a number of the desired value and click submit episode example.

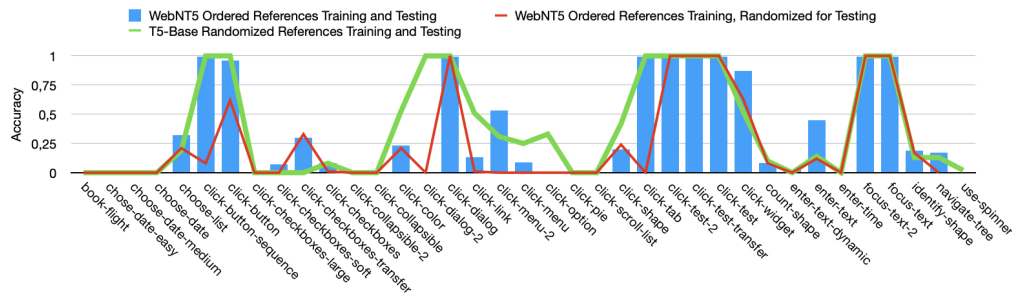


Figure 11: Comparison and effects of fine-tuning with ordered references on a randomized reference test set, and when directly fine-tuned with randomized references.

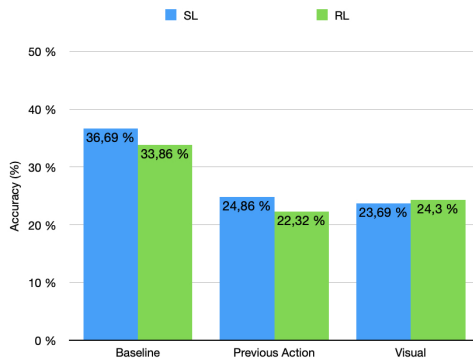


Figure 12: Ablation study of CC-Net5 after its initial SL phase and final RL one.

4.4 Performance of Combining T5 and CC-Net

The combination of T5 and CC-Net was explored in two phases: supervised learning (SL) and reinforcement learning (RL), with ablation studies conducted in both.

4.4.1 Results of Supervised Learning Phase. The initial SL phase achieved 36.69% accuracy, lower than the best T5-large model observed in Table 1. Ablation studies revealed that the model was heavily dependent on the T5-large model, with complete failure when T5’s output was removed. Interestingly, the removal of visual

inputs sometimes improved performance as seen in Figure 14a, indicating a complex interaction between modalities. Overall, the results showed that the model learned a multimodal representation but was mainly reliant on the T5-large model.

4.4.2 Results of Reinforcement Learning Phase. The RL phase resulted in a further drop in accuracy to 33.86% observed in Table 1. Ablation studies confirmed the model’s continued dependence on the T5-large model as seen in Figures 12 for average accuracy and 14b for specific tasks. Several factors may have contributed to this decline, including the model’s inherent complexity, sensitivity to hyperparameter tuning presented in Table 2, the nature of the RL environment, and possible covariate shift between SL and RL phases. These challenges highlight the intricacies of RL training and the need for careful analysis and tuning.

4.5 Benchmarking Results

The model was benchmarked on a subset of Miniwob tasks, focusing on 40 out of 80 available tasks. The results showed competitive performance with previous supervised learning methods while using less training data. A key finding was the tendency of previous models to memorize rather than understand the distribution of target elements. This work’s benchmarking results, though sometimes lower, offer a more robust and reliable assessment of performance.

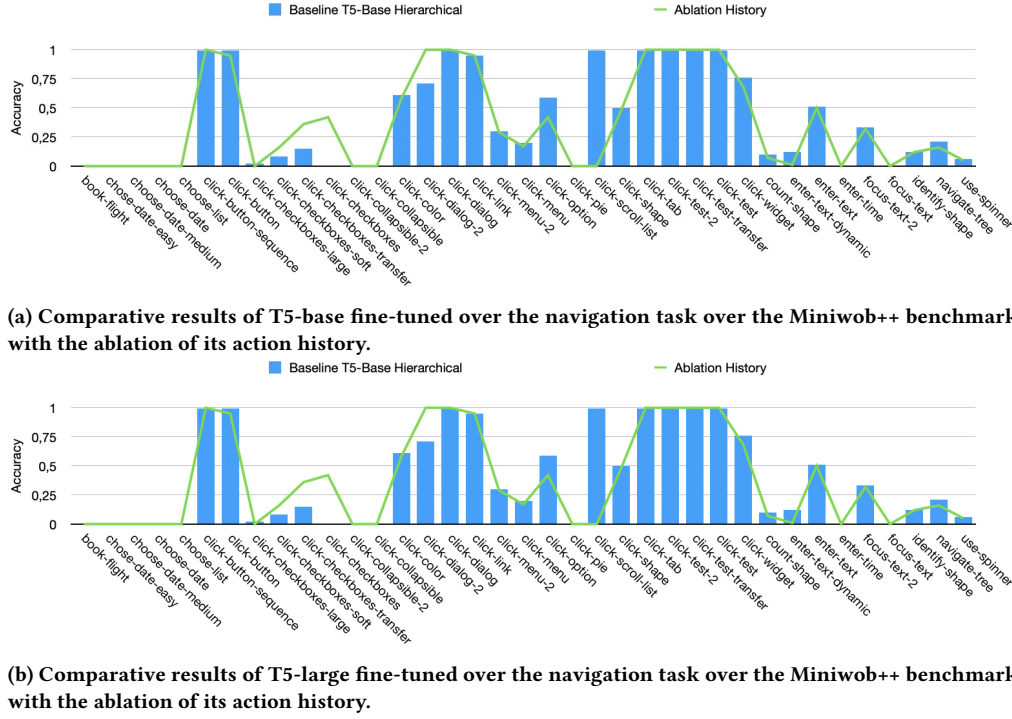


Figure 13: Comparative results of the different T5-only models over the navigation task with an ablation study of their inputs over the Miniwob++ benchmark.

5 DISCUSSION AND LIMITATIONS

Models trained solely on Miniwob are proficient but lack transferability to diverse web tasks. Pretrained models like T5 have limitations in input context and are prone to overfitting. Our approach in Miniwob++ establishes a more realistic baseline, albeit sometimes lower than prior works [10, 33], but we showed it is more grounded to its environment.

Performance evaluation needs to be multifaceted, incorporating metrics beyond task accuracy, such as generalization abilities. Models remain data-intensive and susceptible to memorization; alternative approaches like RLHF should be explored [10, 33]. Despite their limitations, large pre-trained models still perform best, but intermediate-sized models remain under-explored. Small models excel in task-constrained settings but struggle with complexity, while large models offer better generalization but are often overqualified for tasks they solve [14, 18]. The integration of multimodal approaches, like ours with T5, reveals tokenization discrepancies that need optimization [14, 18], and training improvements may include randomizing DOM elements and using ablated inputs to focus on content rather than pattern memorization.

Web navigation automation involves critical ethical and legal aspects that demand attention. Ensuring user privacy is paramount, requiring secure data handling and compliance with regulations like GDPR [11]. The growing capability of large language models to mimic humans raises ethical concerns about impersonation [5], outpacing traditional identifiers like the Turing test [4]. Finally,

clear accountability must be established to navigate complex liability issues, including potential violations of copyright laws and data protection policies. These challenges highlight the need for robust regulations and well-defined licensing agreements.

6 CONCLUSION

Automating web navigation tasks offers many benefits but also presents significant challenges.

Our behavioral cloning and hierarchical planning models achieved a top accuracy of 43.58% on the Miniwob++ benchmark, setting a more grounded performance baseline by mitigating overfitting tendencies commonly seen in large language models (LLMs). While the fine-tuned multimodal model excelled in behavioral cloning, its reinforcement learning phase was hindered by covariate shift due to architectural limitations. We highlight the need for further exploration in multimodal architecture design and the potential of fine-tuned LLMs.

This work emphasizes the importance of understanding the limitations of LLMs over web navigation tasks, exploring optimal architectures, and considering ethical implications. While large models excel in few-shot scenarios, smaller models are efficient for known environments. The exploration of intermediate sizes and multimodal models offers promising opportunities. Pre-processing techniques can alleviate some issues, but further efforts are needed for more capable models. Ethical considerations, including misuse, impersonation, and accountability, must be addressed to ensure the

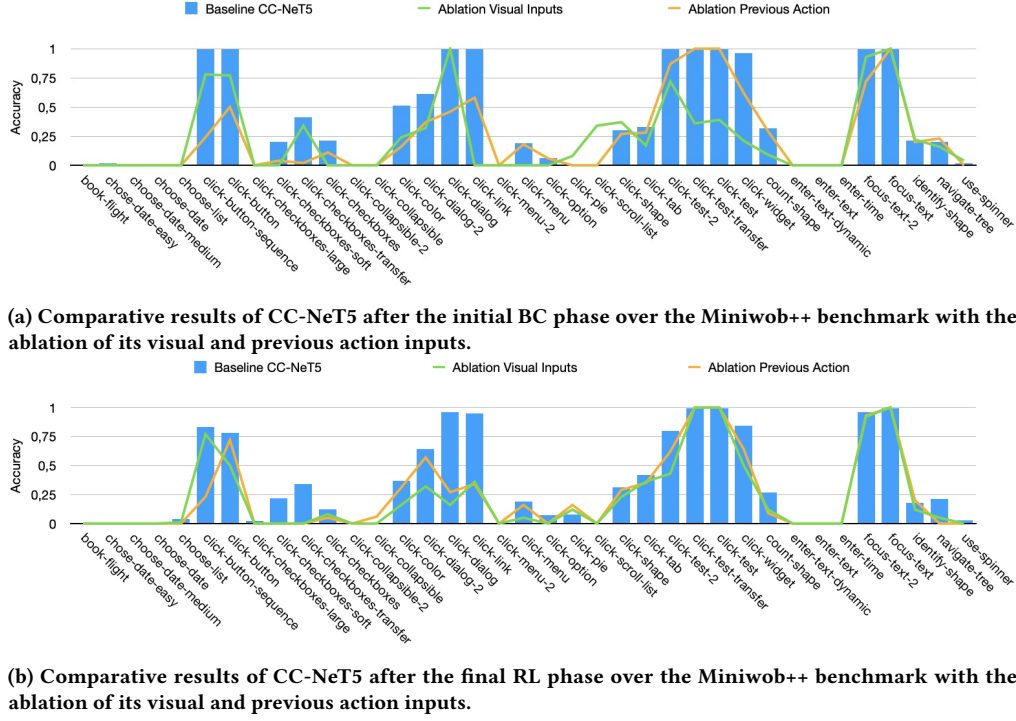


Figure 14: Comparative results of the CC-NeT5 architecture over the navigation task with an ablation study of their inputs over the Miniwob++ benchmark for both supervised and reinforcement learning.

safety and integrity of the web. This work contributes by highlighting these challenges and proposing techniques to overcome them, paving the way for future advancements in the field.

7 APPENDIX

Table 1: Average Accuracy of the Different Models, Fine-Tuned T5 and Combined with CC-Net.

Model Name	Average Accuracy
T5-Base Hierarchical	39.77%
T5-Base Hierarchical No-History	38.02%
T5-Large Hierarchical	43.49%
T5-Large Hierarchical No-History	43.58%
WebNT5-Base	37.94%
WebNT5-Base No-History	35.99%
WebNT5-Base Ordered Refs	35.25%
WebNT5-Base Randomized Refs Test	24.19%
CC-NeT5 Hierarchical (SL)	36.69%
CC-NeT5 Hierarchical (SL+RL)	33.86%

REFERENCES

- [1] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. 2020. An Optimistic Perspective on Offline Reinforcement Learning. arXiv:1907.04543 [cs.LG]
- [2] Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. HTML: Hyper-Text Pre-Training and Prompting of Language Models. arXiv:2107.06955 [cs.CL]
- [3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A Closer Look at Memorization in Deep Networks. arXiv:1706.05394 [stat.ML]
- [4] Aladdin Ayesh. 2019. Turing Test Revisited: A Framework for an Alternative. arXiv:1906.11068 [cs.AI]
- [5] Jean Baudrillard. 1981. *Simulacres et Simulations*. Galilée.
- [6] S.R.K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement Learning for Mapping Instructions to Actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 82–90. <https://aclanthology.org/P09-1010>
- [7] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. arXiv:1606.01540 [cs.LG]

Table 2: Hyper-parameters used in V-MPO during the RL 'online' phase of our training, inspired by CC-Net training parameters.

Parameter	Value
Optimizer	Adam [17]
Learning rate	1e-4
Adam b1 parameter	0.9
Adam b2 parameter	0.999
Weight decay (biases excluded)	1e-1
VMPO α	0.1
VMPO η	0.2
Agent discount γ	0.9
Batch size SL	120
Trajectory unroll length	64
Target-network update period T	5
Maximum number of steps per episode	10

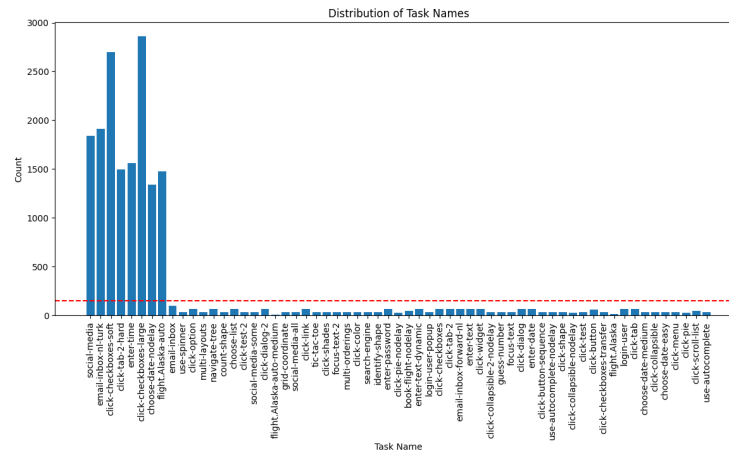


Figure 15: Distribution of task names from the original Miniwob dataset. We can observe that some tasks are over-represented, therefore we sample a maximum of 150 episodes for the examples exceeding that number. This leads to an average amount of 56.29 of episodes per task in the dataset.

- [8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]
- [9] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. arXiv:2202.07646 [cs.LG]
- [10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0dad503c91f91df2400dc4e49-Paper.pdf
- [11] European Parliament and Council of the European Union. [n. d.]. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. <https://data.europa.eu/eli/reg/2016/679/oj>
- [12] Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2023. Understanding HTML with Large Language Models. arXiv:2210.03945 [cs.LG]
- [13] Zecheng He, Srinivas Sunkara, Xiaoxue Zang, Ying Xu, Lijuan Liu, Nevan Wichers, Gabriel Schubiner, Ruby Lee, Jindong Chen, and Blaise Agüera y Arcas. 2021. ActionBERT: Leveraging User Actions for Semantic Understanding of User Interfaces. arXiv:2012.12350 [cs.CL]
- [14] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556 [cs.CL]
- [15] Peter C Humphreys, David Raposo, Toby Pohlen, Gregory Thornton, Rachita Chhaparia, Alistair Muldal, Josh Abramson, Petko Georgiev, Alec Goldin, Adam Santoro, and Timothy Lillicrap. 2022. A data-driven approach for learning to control computers. arXiv:2202.08137 [cs.LG]
- [16] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language Models can Solve Computer Tasks. arXiv:2303.17491 [cs.CL]
- [17] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]
- [18] Ian LeCun. [n. d.]. Do large language models need sensory grounding for meaning and understanding?
- [19] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://www.aclweb.org/anthology/W04-1013>
- [20] Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. 2018. Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration. In *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1802.08802>
- [21] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332 [cs.CL]
- [22] OpenAI. 2016. Universe. <https://github.com/openai/universe>
- [23] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [24] Panupong Pasupat, Tian-Shun Jiang, Evan Liu, Kelvin Guu, and Percy Liang. 2018. Mapping natural language commands to web elements. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4970–4976. <https://doi.org/10.18653/v1/D18-1540>
- [25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]
- [26] Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 41–45. <https://aclanthology.org/E17-2007>
- [27] Julian Schrittwieser, Thomas Hubert, Amol Mandhane, Mohammadamin Barekatain, Ioannis Antonoglou, and David Silver. 2021. Online and Offline Reinforcement Learning by Planning with a Learned Model. arXiv:2104.06294 [cs.LG]
- [28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG]
- [29] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of Bits: An Open-Domain Platform for Web-Based Agents. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3135–3144. <https://proceedings.mlr.press/v70/shi17a.html>
- [30] H. Francis Song, Abbas Abdomaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W. Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Dan Belov, Martin Riedmiller, and Matthew M. Botvinick. 2019. V-MPO: On-Policy Maximum a Posteriori Policy Optimization for Discrete and Continuous Control. arXiv:1909.12238 [cs.AI]
- [31] Adam Vogel and Daniel Jurafsky. 2010. Learning to Follow Navigational Directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 806–814. <https://aclanthology.org/P10-1083>
- [32] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2023. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. arXiv:2207.01206 [cs.CL]
- [33] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs.CL]