Πολυτεχνική Σχολή
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Διπλωματική Εργασία

---

# Τεχνικές Μηχανικής Μάθησης και Δέντρων Επιθημάτων για πρόβλεψη και σύσταση σελίδων στο διαδίκτυο

---

Κανελλόπουλος Παύλος
1084575

*Επιβλέπων*
Μακρής Χρήστος
Αναπληρωτής Καθηγητής

Πάτρα, 2026

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

**Τριμελής Επιτροπή Αξιολόγησης:**

Χρήστος Μακρής, Αναπληρωτής Καθηγητής

Αριστείδης Ηλίας, ΕΔΙΠ

Σπύρος Σιούτας, Καθηγητής

© Copyright συγγραφής Κανελλόπουλος Παύλος 2026

© Copyright θέματος Κανελλόπουλος Παύλος, Μακρής Χρήστος

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών & Πληροφορικής του Πανεπιστημίου Πατρών δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

# Περίληψη

Ένα από τα πιο καίρια και απαιτητικά προβλήματα που κλίνονται να διερευνήσουν και να επιλύσουν οι ερευνητές σήμερα είναι η αυτόνομη πλοήγηση agents σε ιστοσελίδες. Αυτό διότι οι agents βασιζόμενοι σε γλωσσικά μοντέλα καλούνται πλέον να λάβουν αποφάσεις επιλέγοντας την επόμενη σωστή ενέργεια έχοντας ως είσοδο το περιεχόμενο μιας ιστοσελίδας (observation), το οποίο είναι συχνά εκτενές, θορυβώδες και ημιδομημένο. Το LCoW εισήγαγε μια σύγχρονη και ισχυρή μέθοδο προσέγγισης στο πρόβλημα αυτό. Εκπαιδεύοντας ένα εξειδικευμένο module εννοιολογικής πλαισίωσης (contextualization module) επιχειρούν να μετασχηματίσουν το εκάστοτε observation σε μια πιο καθαρή και αξιοποιήσιμη μορφή πριν την λήψη της απόφασης.

Στόχος συνεπώς της παρούσα εργασίας είναι να διερευνήσουμε μια εναλλακτική τοποθέτηση της παρέμβασης οπότε αντί να ενισχύσουμε το στάδιο μετασχηματισμού της σελίδας, επεμβαίνουμε στο ίδιο το στάδιο λήψης της απόφασης. Για τον σκοπό αυτό, προτείνουμε έναν υβριδικό μηχανισμό που συνδυάζει ένα μικρό γλωσσικό μοντέλο με μια συμβολική, μη-παραμετρική δομή δεδομένων, ένα Γενικευμένο Δέντρο Επιθημάτων (Generalized Suffix Tree, GST), το οποίο κωδικοποιεί ιστορικά μοτίβα ενεργειών από προηγούμενες πλοηγήσεις τα οποία αξιοποιούνται στη συνέχεια ως μια καθοδήγηση την στιγμή της λήψης της απόφασης.

Στην προτεινόμενη αυτή υβριδική αρχιτεκτονική, το GST παρέχει σε κάθε βήμα στατιστικές υποδείξεις (hints) βασισμένες στο ιστορικό, ενώ το γλωσσικό μοντέλο διατηρεί πάντα την τελική ευθύνη της απόφασης. Προσαρμόζουμε την μεθοδολογία fine-tuning του LCoW έτσι ώστε: εκπαιδεύουμε το ίδιο μοντέλο απόφασης να αξιοποιεί τις υποδείξεις του GST. Για κάθε βήμα μιας ακολουθίας ενεργειών (trajectory), το GST παράγει πολλαπλά υποψήφια σύνολα hints. Καθένα από αυτά βαθμολογείται με βάση το αν οδηγεί επιτυχώς το μοντέλο στη σωστή ενέργεια και το επικρατέστερο σύνολο διατηρείται για την εκπαίδευση του μοντέλου. Με τον τρόπο αυτόν, το μοντέλο μαθαίνει σταδιακά να αξιοποιεί τα ιστορικά θεμελιωμένα hints κατά τη λήψη της απόφασης.

Η πειραματική μελέτη στο περιβάλλον WebShop ανέδειξε μια σειρά ευρημάτων με ευρύτερη σημασία. Πρώτον, οι υποδείξεις του Δέντρου είναι απαραίτητες για το μικρό γλωσσικό μοντέλο: χωρίς αυτές, η ακρίβεια του μη fine-tuned μοντέλου πέφτει από 78,8% σε 21,2%, κυρίως επειδή αδυνατεί να παραγάγει έγκυρη ενέργεια. Δεύτερον, το μη fine-tuned μοντέλο δεν αξιοποιεί αξιόπιστα τις υποδείξεις και η αξιοποίησή τους απαιτεί fine-tuning. Μέσω ελεγχόμενης πειραματικής σύγκρισης και ανάλυσης bootstrap σε επίπεδο trajectory, το fine-tuning βελτιώνει την ακρίβεια κατά +17,8 ποσοστιαίες μονάδες σε επίπεδο βήματος και +19,4 μονάδες σε επίπεδο ολόκληρου trajectory, με

μείωση των σφαλμάτων κατά περίπου 50%, όταν οι υποδείξεις περιέχουν τη σωστή ενέργεια. Τέλος, αναδεικνύεται πως το Δέντρο Επιθημάτων δεν ανταγωνίζεται το γλωσσικό μοντέλο αλλά το συμπληρώνει, παρέχοντας ορίσματα όπως κωδικούς προϊόντων, που το μοντέλο αδυνατεί να ανακτήσει από μόνο του.

Ακόμα, καθοριστικό μέρος της μελέτης υπήρξε η διασφάλιση της εγκυρότητας των αποτελεσμάτων. Κατά τη διερεύνηση εντοπίστηκαν και αντιμετωπίστηκαν τέσσερα ζητήματα που αλλοίωναν τις αρχικές μετρήσεις. Συνεπώς, εντοπίστηκε data leak στο baseline και μια αστοχία στην αντιστοίχιση των hints, τα οποία διορθώθηκαν στην υλοποίηση, καθώς και δύο ζητήματα εγγενή στα δεδομένα του benchmark τα οποία αφορούν την ένωση επαναλαμβανόμενων επεισοδίων του ίδιου task (από τις διαδοχικές επαναλήψεις συλλογής) που προκαλούσε διαρροή εντός συμφραζομένων, αλλά και η χρήση μερικώς αποτυχημένων trajectories ως ετικετών αξιολόγησης. Η αντιμετώπιση όλων με συνεπή, αυστηρή μεθοδολογία (και επαλήθευση της βασικής δομής μέσω εξαντλητικής σύγκρισης) διασφαλίζει ότι τα αναφερόμενα αποτελέσματα είναι έγκυρα, και αναδεικνύει πώς τέτοιες παγίδες αξιολόγησης μπορούν να οδηγήσουν σε φαινομενικά ισχυρά αλλά μη αξιόπιστα ευρήματα.

Ακόμα, σημειώνεται ότι το καθαρό σύνολο αξιολόγησης είναι περιορισμένο (n=45 στην κύρια ανάλυση, 113 στην επεκταμένη), συνέπεια της ακριβούς συμβολικής αντιστοίχισης και της απουσίας διαδραστικής εκτέλεσης και γι' αυτό τα ευρήματα της αρχικής εκπαίδευσης παρουσιάζονται ως ενδεικτικά, με ευρέα διαστήματα εμπιστοσύνης, ενώ οι μεγαλύτερες διαφορές του πειράματος ελέγχου της υπόθεσης είναι στατιστικά σημαντικές.

Η επεκτεταμένη αξιολόγηση σε 113 σημεία ανέδειξε αρχικά έναν σημαντικό περιορισμό: όταν η σωστή ενέργεια απουσιάζει από τις υποδείξεις (OOD), το fine-tuned μοντέλο αντιγράφει το κορυφαίο hint και πέφτει στο 27,9%, ενώ το μη fine-tuned διαβάζει στόχο και σελίδα και φτάνει το 88,2%. Η ανάλυση εντόπισε την πιθανή αιτία στον τρόπο κατασκευής των δεδομένων εκπαίδευσης: επειδή το Δέντρο περιείχε και τα ίδια τα επεισόδια εκπαίδευσης, η σωστή ενέργεια βρισκόταν πάντα μέσα στις υποδείξεις. Η υπόθεση ελέγχθηκε με προκαταχωρισμένο πείραμα, στο οποίο οι υποδείξεις εκπαίδευσης κατασκευάζονται με αποκλεισμό του ίδιου task (leave-one-task-out). Το νέο μοντέλο φτάνει 92,6% στις καταστάσεις OOD (+64,7 μονάδες, 95% CI [+52,9, +76,5]), αυξάνει το όφελος όταν οι υποδείξεις είναι σωστές (91,1% έναντι 64,4% του μη fine-tuned, 95% CI της διαφοράς [+11,6, +41,9]) και πετυχαίνει 92,0% στο πλήρες σύνολο, την υψηλότερη επίδοση όλων των διαμορφώσεων. Το αποτέλεσμα επιβεβαιώθηκε με τρία διαφορετικά seeds εκπαίδευσης.

# Abstract

One of the most critical and challenging problems that researchers are currently seeking to investigate and solve is the autonomous navigation of agents on websites. This is because agents based on language models are now required to make decisions by selecting the next appropriate action based on the content of a web page (observation), which is often extensive, noisy, and semi-structured. LCoW introduced a modern and powerful approach to this problem. By training a specialized contextualization module, they aim to transform each observation into a clearer and more usable form before a decision is made.

The goal of this paper, therefore, is to explore an alternative approach to intervention: rather than enhancing the page transformation stage, we intervene in the decision-making stage itself. To this end, we propose a hybrid mechanism that combines a small language model with a symbolic, non-parametric data structure, a Generalized Suffix Tree (GST), which encodes historical patterns of actions from previous navigations that are then utilized as guidance at the moment of decision-making.

In this proposed hybrid architecture, the GST provides history-based statistical hints at each step, while the language model always retains ultimate responsibility for the decision. We adapt the LCoW fine-tuning methodology so that we train the same decision model to utilize the GST's hints. For each step of a trajectory, the GST generates multiple candidate sets of hints. Each of these is scored based on whether it successfully guides the model to the correct action, and the most promising set is retained for model training. In this way, the model gradually learns to utilize historically grounded hints when making decisions.

Our experimental study in the WebShop environment revealed a series of findings with broader implications. First, the tree's hints are essential for the small language model: without them, the accuracy of the non-fine-tuned model drops from 78.8% to 21.2%, mainly because it fails to produce a valid action. Second, the non-fine-tuned model does not use the hints reliably, and using them requires fine-tuning. Through a controlled comparison and trajectory-level bootstrap analysis, fine-tuning improves accuracy by +17.8 percentage points at the step level and +19.4 percentage points at the trajectory level, with roughly half the errors, when the hints contain the correct action. Finally, the Suffix Tree does not compete with the language model but complements it, providing arguments such as product codes that the model cannot retrieve on its own.

Furthermore, ensuring the validity of the results was a crucial part of the study. During the investigation, four issues that distorted the initial measurements were identified and addressed. Consequently, a data leak was detected in the baseline and a failure in the matching of hints, both of which were corrected in the implementation, as well as two issues inherent in the benchmark data concerning the merging of repeated instances of the same task (from successive collection iterations) that caused in-context leakage, as

well as the use of partially failed trajectories as evaluation labels. Addressing all of these with a consistent, rigorous methodology (and verification of the underlying structure through exhaustive comparison) ensures that the reported results are valid, and highlights how such evaluation pitfalls can lead to seemingly robust but unreliable findings.

The clean evaluation set is limited (n=45 in the main analysis, 113 in the extended one), a consequence of exact symbolic matching and the absence of interactive execution. The findings of the original training are therefore presented as indicative, with wide confidence intervals, while the larger differences of the hypothesis-testing experiment are statistically significant.

The extended 113-point evaluation first revealed a significant limitation: when the correct action is absent from the hints (OOD), the fine-tuned model copies the top hint and drops to 27.9%, while the non-fine-tuned model reads the goal and the page and reaches 88.2%. The analysis traced the likely cause to how the training data were built: because the tree also contained the training episodes themselves, the correct action was always among the hints. We tested this hypothesis with a pre-registered experiment in which training hints are retrieved with the example's own task excluded (leave-one-task-out). The new model reaches 92.6% on the OOD points (+64.7 points, 95% CI [+52.9, +76.5]), increases the gain when the hints are correct (91.1% vs 64.4% for the non-fine-tuned model, 95% CI of the difference [+11.6, +41.9]) and achieves 92.0% on the full set, the highest of all configurations. The result held across three training seeds.

# Περιεχόμενα

*Περίληψη* ..... *iv*

*Abstract* ..... *vi*

*Λίστα Εικόνων* ..... *x*

*Λίστα Πινάκων* ..... *xi*

*Εισαγωγή* ..... *1*

**1.1 Σημασία του Προβλήματος** ..... **1**

**1.2 Συνεισφορά** ..... **3**

**1.3 Διάρθρωση** ..... **5**

*2 Υπόβαθρο* ..... *6*

**2.1 Μεγάλα Γλωσσικά Μοντέλα και Αυτόνομοι Πράκτορες** ..... **6**

**2.2 Το πρόβλημα της πλοήγησης στον ιστό και το περιβάλλον WebShop** ..... **9**

**2.3 Fine-tuning και τεχνική LoRA** ..... **11**

**2.4 Η μέθοδος LCoW** ..... **13**

**2.5 Γενικευμένο Δέντρο Επιθημάτων (Generalized Suffix Tree, GST)** ..... **15**

*3 Σχετική Βιβλιογραφία* ..... *18*

**3.1 Πράκτορες πλοήγησης Ιστού** ..... **18**

**3.2 Επεξεργασία των observations και Contextualization Module** ..... **19**

**3.3 Συμβολικές και Υβριδικές Προσεγγίσεις** ..... **20**

**3.4 Θέση της παρούσας Εργασίας** ..... **22**

*Αρχιτεκτονική* ..... *23*

**4.1 Επισκόπηση Αρχιτεκτονικής** ..... **23**

**4.2 Κατασκευή του GST** ..... **26**

**4.3 Μηχανισμός ιστορικών υποδείξεων (hints)** ..... **31**

**4.4 Ενσωμάτωση LLM και Always-Decide approach** ..... **34**

**4.5 Σχεδιασμός Δεδομένων προσαρμογής** ..... **38**

**4.6 Σύνοψη Σχεδιαστικών Αποφάσεων** ..... **41**

*Υλοποίηση και Αποτελέσματα* ..... *43*

**5.1 Υπολογιστικό περιβάλλον και αναπαραγωγιμότητα ..... 43**
**5.2 Διαχωρισμός Δεδομένων και Αποτροπή Διαρροής ..... 45**
**5.3 Ρυθμίσεις LoRA ..... 48**
**5.4 Πειραματική Διάταξη ..... 51**
**5.5 Κύρια Αποτελέσματα ..... 54**
**5.6 Ανάλυση Σφαλμάτων ..... 59**
**5.7 Συμπληρωματικά ευρήματα ..... 61**
**5.8 Αξιολόγηση σε καταστάσεις χωρίς ιστορικό προηγούμενο ..... 63**
**5.9 Έλεγχος της υπόθεσης: εκπαίδευση με υποδείξεις που αποκλείουν το ίδιο task ..... 64**
***Συμπεράσματα και Μελλοντικές Προεκτάσεις ..... 68***
**6.1 Σύνοψη Συνεισφορών ..... 68**
**6.2 Περιορισμοί ..... 70**
**6.3 Μελλοντικές Προεκτάσεις ..... 73**

# Λίστα Εικόνων

Εικόνα 2.1: Παράδειγμα δομής ενός Γενικευμένου Δέντρου Επιθημάτων για τις ακολουθίες «ABC» και «ABD»……17

Εικόνα 4.1: Συνολική αρχιτεκτονική του υβριδικού συστήματος GST-LLM…….…..28

Εικόνα 4.2: Πραγματική γραμμή των δεδομένων πριν τον διαχωρισμό σε επεισόδια. Μετά το click[buy now] ξεκινά δεύτερο επεισόδιο του ίδιου task, γι' αυτό τα observations 0 και 1 επανεμφανίζονται.…… ……..29

Εικόνα 4.3: Περιεχόμενο ενός πραγματικού κόμβου του Δέντρου……32

Εικόνα 4.4: Κατανομή κόμβων ανά βάθος…….34

Εικόνα 4.5: Οι πέντε κόμβοι του πρώτου επιπέδου με το υψηλότερο visit_count…….35

Εικόνα 4.6: Ενδεικτικό παράδειγμα του μηχανισμού υποχώρησης (backoff): διαδοχική βράχυνση του επιθήματος έως την πρώτη έγκυρη αντιστοίχιση (εδώ στο μήκος 1)……..36

Εικόνα 4.7: Σχηματικό διάγραμμα των δύο στρατηγικών δρομολόγησης…….….40

Εικόνα 4.8: Πραγματική προτροπή και εσφαλμένη απόκριση του μη fine-tuned μοντέλου, παρά την ορθή κορυφαία υπόδειξη…….…….41

Εικόνα 4.9: Παραγωγή παραδείγματος εκπαίδευσης με βαθμολόγηση ανταμοιβής και μηχανισμό fallback για τα βήματα μηδενικής ανταμοιβής……..43

Εικόνα 5.1: Διαχωρισμός σε επίπεδο task. Η ένωση των ids τέμνεται σε ξένα σύνολα ώστε κανένα task να μην εμφανίζεται και στις δύο πλευρές…….....52

Εικόνα 5.2: Ακρίβεια επιπέδου step ανά διαμόρφωση, με 95% διαστήματα εμπιστοσύνης (Cluster Bootstrap ανά trajectory, 10.000 επαναλήψεις)…….60

Εικόνα 5.3: Κατανομή των 36 trajectories κατά πλήθος λαθών, πριν και μετά το fine-tuning (real observation)…….62

Εικόνα 5.4: Οι τέσσερις επιδράσεις με τα κατά ζεύγη 95% διαστήματα εμπιστοσύνης. Τα ευρέα διαστήματα αντανακλούν το μικρό μέγεθος δείγματος (n=45)· καμία επίδραση δεν αποκλείει τυπικά το μηδέν.……...……64

Εικόνα 5.5: Κατανομή Λαθών ανά Κατηγορία, πριν και μετά το fine-tuning (real observations)……67

# Λίστα Πινάκων

Πίνακας 4.1: Στατιστικά του Δέντρου……33

Πίνακας 4.2: Οι δύο παραλλαγές αναζήτησης με υποχώρηση…….38

Πίνακας 4.3: Σύνθεση του συνόλου προσαρμογής (923 παραδείγματα)…….46

Πίνακας 4.4: Η λογική του Συστήματος……48

Πίνακας 5.1: Σύνοψη των συνόλων δεδομένων……54

Πίνακας 5.2: Πλήρης διαμόρφωση της προσαρμογής LoRA……56

Πίνακας 5.3: Οι πέντε πειραματικές διαμορφώσεις……..59

Πίνακας 5.4: Ακρίβεια επιπέδου step ανά διαμόρφωση (n=45 σημεία, 36 trajectories seed=42)…….……..63

Πίνακας 5.5: Οι τέσσερις επιδράσεις με τα κατά ζεύγη διαστήματα εμπιστοσύνης (τα διαστήματα εμπιστοσύνης είναι ευρέα λόγω του μεγέθους των δειγμάτων (n=45))……...……...63

Πίνακας 5.6: Κατανομή λαθών ανά κατηγορία για τις δύο διαμορφώσεις με real observations.....66

Πίνακας 5.7: Ακρίβεια με 95% CI……..……71

Πίνακας 5.8: Σύνθεση των δύο συνόλων προσαρμογής……73

Πίνακας 5.9: Ακρίβεια με 95% CI, με τις δύο διαμορφώσεις της εκπαίδευσης leave-one-task-out (LOTO)……..……73

Πίνακας 5.10: Ακρίβεια του fine-tuned LOTO με real observation σε τρία seeds εκπαίδευσης……..……74

# 1

# Εισαγωγή

## 1.1 Σημασία του Προβλήματος

Με την ραγδαία εξέλιξη της τεχνολογίας τα τελευταία χρόνια, ειδικότερα στον τομέα των μεγάλων γλωσσικών μοντέλων (Large Language Models , LLMs) έχει δημιουργηθεί η ανάγκη για ολοένα και πιο αυτόνομα συστήματα, ικανά όχι απλώς να παράγουν κείμενο που να προσομοιώνει όσο καλύτερα γίνεται τον άνθρωπο, αλλά να αναλαμβάνουν να φέρουν εις πέρας πιο σύνθετες εργασίες όπως η αυτόνομη πλοήγηση σε ιστοσελίδες (web navigation) χωρίς συνεχή ανθρώπινη καθοδήγηση. Συνεπώς, το ενδιαφέρον έχει στραφεί προς την δημιουργία αυτόνομων agents, οι οποίοι αξιοποιώντας τις τεράστιες δυνατότητες των LLMs να μπορούν να αλληλοεπιδρούν με το περιβάλλον τους ολοκληρώνοντας ενέργειες με σκοπό την ολοκλήρωση ενός στόχου που τους έχει τεθεί από έναν χρήστη σε φυσική γλώσσα. Η πρακτική αξία αυτών των αυτόνομων συστημάτων είναι ιδιαίτερα μεγάλη, καθώς στο μέλλον θα μπορέσουν να αυτοματοποιήσουν ένα ευρύ φάσμα καθημερινών εργασιών , από την αναζήτηση και σύγκριση προϊόντων στο διαδίκτυο έως την συμπλήρωση φορμών και την διεκπεραίωση συναλλαγών, κατά συνέπεια εξοικονομώντας σημαντικό χρόνο και προσπάθεια για τον χρήστη.

Η διεκπεραίωση, ωστόσο, ενός τέτοιου εγχειρήματος προαπαιτεί και την επίλυση ορισμένων ιδιαίτερων προκλήσεων. Οι agents δεν λαμβάνουν μια δομημένη περιγραφή της κατάστασης στην οποία βρίσκονται και κατά συνέπεια καλούνται να λάβουν αποφάσεις επιλέγοντας την επόμενη σωστή ενέργεια έχοντας ως είσοδο το ακατέργαστο περιεχόμενο μιας ιστοσελίδας (observation), το οποίο είναι κατά κανόνα εκτενές, θορυβώδες και ημιδομημένο. Επιπλέον, η πλοήγηση είναι μια ακολουθιακή διαδικασία

όπου κάθε ενέργεια που κάνει ο agent μεταβάλλει την κατάσταση του περιβάλλοντος στο οποίο βρίσκεται και επηρεάζει τις επιλογές που θα είναι διαθέσιμες στην συνέχεια. Ως αποτέλεσμα αυτού, ένα σφάλμα στην διαδικασία σε πρώιμο βήμα μπορεί να υπονομεύσει ολόκληρη την υπόλοιπη πορεία του agent μέσα στο περιβάλλον έχοντας αρνητικό τελικό αποτέλεσμα. Η ανάγκη, επομένως , για ακριβή κατανόηση της κατάστασής του και συνεπώς του περιεχομένου για ορθή λήψη αποφάσεων σε κάθε βήμα καθιστά το πρόβλημα της αυτόνομης πλοήγησης των agents ιδιαιτέρως σύνθετο.

Για την αντιμετώπιση αυτών των δυσκολιών, οι ερευνητές έχουν προτείνει ποικίλες μεθόδους, ορισμένες εκ των οποίων επικεντρώνονται στην βελτίωση της ποιότητας της πληροφορίας που φτάνει στο γλωσσικό μοντέλο πριν λάβει τις αποφάσεις. Ένα τέτοιο παράδειγμα αποτελεί το LCoW, το οποίο εισάγει ένα εξειδικευμένο module εννοιολογικής πλαισίωσης (contextualization module). Ένα ξεχωριστό νευρωνικό δίκτυο δηλαδή που εκπαιδεύεται ώστε να μετασχηματίζει το ακατέργαστο observation σε μια καθαρότερη, πιο συνεκτική μορφή κρατώντας μόνο την απαραίτητη για τον agent πληροφορία, πριν αυτή δοθεί στο μοντέλο που λαμβάνει την τελική απόφαση. Η προσέγγιση αυτή έχει αποδεδειγμένα δείξει πως βελτιώνει την διαδικασία λήψης απόφασης του agent καταδεικνύοντας πως η κατάλληλη προεπεξεργασία του observation μπορεί να βελτιώσει σημαντικά την επίδοση του agent.

Παρά τα αισιόδοξα θετικά αποτελέσματα της μεθόδου αυτής, η στρατηγική της εστιάζει αποκλειστικά στο προπαρασκευαστικό στάδιο, δηλαδή στο πως θα αναδιατυπωθεί η είσοδος του agent, αφήνοντας ανεκμετάλλευτη μια εναλλακτική δυνατότητα που είναι η άμεση ενίσχυση του ίδιου του σταδίου της λήψης απόφασης. Σε αυτό ακριβώς το σημείο είναι που τοποθετείται η παρούσα εργασία διερευνώντας κατά πόσο η αξιοποίηση ιστορικών μοτίβων από προηγούμενες πλοηγήσεις, μέσω μιας συμβολικής δομής δεδομένων, μπορεί να καθοδηγήσει τον agent κατά την λήψη της απόφασης, προσφέροντας μια αποδοτικότερη εναλλακτική.

Η πρακτική σημασία της αυτοματοποιημένης πλοήγησης στον ιστό, δηλαδή, εκτείνεται πολύ περαιτέρω από το θεωρητικό ενδιαφέρον. Ένα μεγάλο μέρος της καθημερινής ψηφιακής δραστηριότητας, από τη σύγκριση προϊόντων δηλαδή και τη συμπλήρωση φορμών έως τη διαχείριση λογαριασμών και την αναζήτηση πληροφοριών, διαμεσολαβείται από γραφικές διεπαφές σχεδιασμένες για ανθρώπους. Ένας agent ικανός να αντιλαμβάνεται και να χειρίζεται αξιόπιστα τέτοιες διεπαφές θα μπορούσε να αυτοματοποιήσει επαναλαμβανόμενες εργασίες, να ενισχύσει την προσβασιμότητα για χρήστες με δυσκολίες, και να λειτουργήσει ως δομικό στοιχείο ευρύτερων συστημάτων τεχνητής νοημοσύνης που δρουν αυτόνομα. Ωστόσο, παρά την εντυπωσιακή πρόοδο των γλωσσικών μοντέλων, η μετάβαση από επιδείξεις ελεγχόμενων συνθηκών σε αξιόπιστη λειτουργία υπό πραγματικές συνθήκες παραμένει ανοιχτή πρόκληση. Η αναξιοπιστία στη λήψη διαδοχικών αποφάσεων, όπου ένα μόνο εσφαλμένο βήμα μπορεί να ακυρώσει ολόκληρη την εργασία, αποτελεί ένα από τα κεντρικά εμπόδια και ακριβώς σε αυτό το ζήτημα της αξιοπιστίας εστιάζει η παρούσα εργασία.

## *1.2 Συνεισφορά*

Όπως αναφέρθηκε και προηγουμένως, η παρούσα εργασία αφιερώνει την μελέτη της στο στάδιο της λήψης της απόφασης ενός agent και διερευνά κατά πόσο μια συμβολική αναπαράσταση των ιστορικών μοτίβων από προηγούμενες πλοηγήσεις μπορεί να ενισχύσει την ικανότητα ενός μικρού γλωσσικού μοντέλου να επιλέγει την σωστή ενέργεια δοθείσης ενός στόχου. Οι επιμέρους συνεισφορές της εργασίας συνοψίζονται στα ακόλουθα σημεία.

**Υβριδική αρχιτεκτονική συμβολικού και νευρωνικού συστήματος.**

Προτείνουμε μια αρχιτεκτονική στην οποία ένα Γενικευμένο Δέντρο Επιθημάτων (Generalized Suffix Tree, GST) συνδυάζεται με ένα μικρό γλωσσικό μοντέλο για την ενίσχυση της βεβαιότητας στην λήψη της απόφασης. Το GST κωδικοποιεί τα ιστορικά μοτίβα ενεργειών από προηγούμενες πλοηγήσεις και παρέχει σε κάθε βήμα στατιστικές υποδείξεις προς το LLM, το οποίο όπως είπαμε διατηρεί πάντα την τελική ευθύνη της απόφασης. Σε αντίθεση με τις αμιγώς νευρωνικές προσεγγίσεις όπως είναι η εκπαίδευση του contextualization module, η συμβολική αυτή συνιστώσα είναι μη παραμετρική και υπολογιστικά ελαφριά, καθώς δεν απαιτεί κανένα πρόσθετο πέρασμα από νευρωνικό δίκτυο.

**Πειραματική τεκμηρίωση της αναγκαιότητας του fine-tuning.**

Μέσα από ελεγχόμενη σύγκριση του ίδιου μοντέλου με και χωρίς fine-tuning, φάνηκε πως το μη fine-tuned μοντέλο δεν αξιοποιεί αξιόπιστα τις υποδείξεις του GST, ακόμη και όταν η κορυφαία υπόδειξη είναι η σωστή ενέργεια και ότι απαιτείται στοχευμένο fine-tuning ώστε το μοντέλο να μάθει να σταθμίζει τις υποδείξεις σε συνδυασμό με την τρέχουσα κατάσταση.

**Απομόνωση της καθαρής επίδρασης του fine-tuning.**

Με στόχο να ελέγξουμε την καθαρή βελτίωση η οποία αποδίδεται στο fine-tuning, σχεδιάσαμε ορισμένα πειράματα που απομονώνουν την επίδραση του fine-tuning από άλλους συγχυτικούς παράγοντες, ιδιαιτέρως από την επίδραση της αρχιτεκτονικής επιλογής για την λήψη αποφάσεων. Τα αποτελέσματα εμφάνισαν μια θετική επίδραση της τάξεως των +17,8 ποσοστιαίων μονάδων σε επίπεδο βήματος, συνεπή σε πολλαπλές ανεξάρτητες μετρήσεις.

**Στατιστική τεκμηρίωση, διασφάλιση εγκυρότητας και αυστηρότερη αξιολόγηση.** Πέραν της επίδοσης σε επίπεδο βήματος, το σύστημα αξιολογήθηκε με το αυστηρότερο κριτήριο της ακριβούς αντιστοίχισης σε επίπεδο ολόκληρης trajectory, όπου η επίδραση του fine-tuning παρέμεινε θετική (+19,4 ποσοστιαίες μονάδες), με μείωση των συνολικών σφαλμάτων κατά περίπου 50%. Όλα τα μεγέθη συνοδεύονται από διαστήματα εμπιστοσύνης υπολογισμένα με bootstrap σε επίπεδο trajectory. Κρίσιμο μέρος της συνεισφοράς αυτής υπήρξε η διασφάλιση της εγκυρότητας, αφού κατά τη διερεύνηση

εντοπίστηκαν και αντιμετωπίστηκαν τέσσερα ζητήματα που αλλοίωναν τις μετρήσεις, τα οποία αφορούν: δύο της υλοποίησης (διαρροή στο baseline, αστοχία στην αντιστοίχιση hints) και δύο εγγενή στα δεδομένα του benchmark (διαρροή εντός συμφραζομένων από επαναλαμβανόμενα επεισόδια, χρήση μερικώς αποτυχημένων trajectories ως ετικετών). Η αντιμετώπισή τους με συνεπή μεθοδολογία, και η επαλήθευση της βασικής δομής μέσω εξαντλητικής σύγκρισης, εξασφαλίζει την αξιοπιστία των αναφερόμενων αποτελεσμάτων. Η εργασία τεκμηριώνει επίσης τον περιορισμό της αρχικής εκπαίδευσης: όταν η σωστή ενέργεια λείπει από τις υποδείξεις, το fine-tuned μοντέλο υπολείπεται σημαντικά του μη fine-tuned (Ενότητα 5.8).

**Διάγνωση και διόρθωση της υπερ-εξάρτησης από τις υποδείξεις**

Η αιτία του περιορισμού εντοπίστηκε στην κατασκευή των δεδομένων εκπαίδευσης και ελέγχθηκε με προκαταχωρισμένο πείραμα παρέμβασης, κατά το οποίο οι υποδείξεις εκπαίδευσης αποκλείουν το ίδιο task. Η παρέμβαση εξαλείφει την υπερ-εξάρτηση (από 27,9% σε 92,6% στις καταστάσεις όπου η σωστή ενέργεια λείπει από τις υποδείξεις) και καθιστά το fine-tuned μοντέλο το καλύτερο σε όλα τα υποσύνολα, με στατιστικά σημαντικές διαφορές και σταθερά αποτελέσματα σε τρία seeds εκπαίδευσης (Ενότητα 5.9).

**Ανάδειξη του συμπληρωματικού ρόλου των δύο συνιστωσών.** Τέλος, αναδεικνύεται πως η συμβολική δομή δεδομένων (GST) και το νευρωνικό δίκτυο (LLM) δεν ανταγωνίζονται το ένα το άλλο αλλά επιτελούν διακριτούς αλληλοσυμπληρούμενους ρόλους με το GST να παρέχει τις έγκυρες, διαθέσιμες επιλογές της σελίδας. Το εύρημα αυτό φωτίζει τη σχέση μεταξύ συμβολικής και νευρωνικής συνιστώσας σε υβριδικά συστήματα πλοήγησης. Ειδικότερα, χωρίς τις υποδείξεις η ακρίβεια του μη fine-tuned μοντέλου πέφτει από 78,8% σε 21,2% (Ενότητα 5.7).

## *1.3 Διάρθρωση*

Η παρούσα εργασία οργανώνεται σε έξι κεφάλαια. Αναλυτικότερα:

Αρχικά, το Κεφάλαιο 1 εισάγει το πρόβλημα της αυτοματοποιημένης πλοήγησης σε περιβάλλοντα Ιστού, τεκμηριώνει τη σημασία του και συνοψίζει τις συνεισφορές τις εργασίας.

Κατόπιν, το Κεφάλαιο 2 παρουσιάζει το απαραίτητο θεωρητικό υπόβαθρο: τα μεγάλα γλωσσικά μοντέλα και τους agents που βασίζονται σε αυτά, το περιβάλλον WebShop και τον τυπικό φορμαλισμό του προβλήματος, τη μέθοδο fine-tuning LoRA, την προσέγγιση LCoW που αποτέλεσε την αφετηρία της εργασίας και τα Γενικευμένα Δέντρα Επιθημάτων ως δομή δεδομένων.

Στη συνέχεια, το Κεφάλαιο 3 ανασκοπεί τη σχετική βιβλιογραφία, τοποθετώντας την εργασία στο ευρύτερο πεδίο, δηλαδή οι agents πλοήγησης στον Ιστό, οι τεχνικές επεξεργασίας observations, οι συμβολικές και υβριδικές προσεγγίσεις, καθώς και η διακριτή θέση της παρούσας εργασίας έναντι όλων αυτών.

Το Κεφάλαιο 4 αναπτύσσει την αρχιτεκτονική του προτεινόμενου συστήματος, την κατασκευή του Δέντρου Επιθημάτων από ιστορικά trajectories, τον μηχανισμό παραγωγής ιστορικών υποδείξεων με υποχώρηση, την ενσωμάτωση του LLM και τη στρατηγική δρομολόγηση always-decide, καθώς και τον σχεδιασμό των δεδομένων προσαρμογής.

Ακόμα, το Κεφάλαιο 5 περιγράφει την υλοποίηση και παρουσιάζει τα πειραματικά αποτελέσματα, δηλαδή το υπολογιστικό περιβάλλον, τον διαχωρισμό των δεδομένων σε επίπεδο task, τις ρυθμίσεις του fine-tuning, την πειραματική διάταξη, τα κύρια αποτελέσματα με την στατιστική τους τεκμηρίωση, την ανάλυση σφαλμάτων, τα συμπληρωματικά ευρήματα, την αξιολόγηση σε καταστάσεις χωρίς ιστορικό προηγούμενο και το πείραμα ελέγχου της υπόθεσης με υποδείξεις που αποκλείουν το ίδιο task.

Τέλος, το Κεφάλαιο 6 συνοψίζει τις συνεισφορές, συζητά τους περιορισμούς της εργασίας και προτείνει κατευθύνσεις για μελλοντική έρευνα.

# 2

# Υπόβαθρο

## *2.1 Μεγάλα Γλωσσικά Μοντέλα και Αυτόνομοι Πράκτορες*

Από τα Γλωσσικά Μοντέλα στα Μεγάλα Γλωσσικά Μοντέλα
Ένα γλωσσικό μοντέλο (Language Model) είναι στην ουσία τους πιθανοτικά μοντέλα, τα οποία μαθαίνουν μέσα από μια σειρά εκπαιδεύσεων να προβλέπουν ποιο θα είναι το επόμενο token σε μια ακολουθία, έχοντας ως δεδομένο τα προηγούμενα (Bengio et al., 2003). Η ικανότητα αυτή μπορεί να φαντάζει απλή, ωστόσο για να είναι αποτελεσματική πρέπει το μοντέλο έμμεσα να μπορεί να κωδικοποιήσει την γραμματική, σημασιολογία ακόμη και να κατανοήσει την πρόθεση του χρήστη που έγραψε το prompt.

Ο όρος Μεγάλα Γλωσσικά Μοντέλα (Large Language Models, LLMs) αποτελεί την φυσική εξέλιξη των γλωσσικών μοντέλων, τα οποία έχουν εκπαιδευτεί σε έναν τεράστιο όγκο κειμένων, έχοντας δισεκατομμύρια παραμέτρους (Brown et al., 2020). Η αρχιτεκτονική που κατέστησε εφικτή αυτήν την κλιμάκωση των μοντέλων είναι ο Transformer (Vaswani et al., 2017), ο οποίος μέσω του μηχανισμού προσοχής (attention), επιτρέπει στο μοντέλο να σταθμίζει δυναμικά το βάρος των διαφορετικών τμημάτων της εισόδου κατά την επεξεργασία κάθε token. Η σχέση ανάμεσα στο μέγεθος του μοντέλου, τον όγκο των δεδομένων και την τελική απόδοση έχει μελετηθεί συστηματικά και περιγράφεται από εμπειρικούς νόμους κλιμάκωσης (scaling laws) (Kaplan et al., 2020). Παράλληλα, καθώς αυξάνεται το μέγεθος των μοντέλων και των δεδομένων εκπαίδευσης, έχει παρατηρηθεί η ανάδυση ορισμένων ικανοτήτων (emergent abilities), οι οποίες απουσίαζαν από μικρότερα μοντέλα (Wei et al., 2022a), όπως η επίλυση προβλημάτων βήμα – βήμα (Wei et al., 2022b) ή η ακολούθηση σύνθετων οδηγιών.

## Από την Πρόβλεψη Κειμένου στην Ακολούθηση Οδηγιών

Ωστόσο, ένα γλωσσικό μοντέλο στην προεκπαιδευμένη (pre-trained) μορφή του, αν και μπορεί να προβλέπει με αποτελεσματικό τρόπο το κείμενο, δεν είναι εξ ορισμού σε θέση να ακολουθεί οδηγίες ή να παράγει ουσιώδεις απαντήσεις για τον χρήστη. Προκειμένου να αποκτήσει αυτή την ικανότητα, υποβάλλεται σε ένα ακόμη στάδιο εκπαίδευσης που ονομάζεται instruction tuning (Wei et al., 2021), κατά τη διάρκεια του οποίου το μοντέλο ανατροφοδοτείται με ζεύγη εντολών και αντίστοιχων απαντήσεων έχοντας ως αποτέλεσμα το μοντέλο να μάθει να ανταποκρίνεται ορθά σε αιτήματα που του διατυπώνονται σε φυσική γλώσσα. Κατόπιν, η συμπεριφορά του μοντέλου βελτιώνεται ακόμη περισσότερο μέσα από μεθόδους ευθυγράμμισης, με πιο γνωστή την ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση (Reinforcement Learning from Human Feedback, RLHF) (Ouyang et al., 2022) και με αυτόν τον τρόπο οι απαντήσεις του μοντέλου αποκτούν ολοένα και πιο ανθρώπινη συμπεριφορά μιμούμενο την ανθρώπινη λογική. Αποτέλεσμα αυτής της διαδικασίας είναι τα λεγόμενα μοντέλα «οδηγιών» (instruct models), όπως είναι αυτά που μελετώνται στην παρούσα εργασία.

Η αρχιτεκτονική που κατέστησε δυνατή αυτή την κλιμάκωση είναι ο μετασχηματιστής (Transformer), ο οποίος εισήγαγε τον μηχανισμό της αυτο-προσοχής (self-attention) ως θεμελιώδη δομική μονάδα (Vaswani et al., 2017). Σε αντίθεση με τις προγενέστερες αναδρομικές αρχιτεκτονικές, που επεξεργάζονταν το κείμενο ακολουθιακά και δυσκολεύονταν να συλλάβουν εξαρτήσεις μεγάλης απόστασης, ο μηχανισμός προσοχής επιτρέπει σε κάθε στοιχείο μιας ακολουθίας να «σταθμίζει» τη σημασία όλων των υπόλοιπων στοιχείων ταυτόχρονα. Πρακτικά, για κάθε λέξη-εισόδου το μοντέλο υπολογίζει πόσο σχετική είναι με κάθε άλλη λέξη του συμφραζομένου και διαμορφώνει την αναπαράστασή της ως σταθμισμένο συνδυασμό των υπολοίπων. Η ιδιότητα αυτή, σε συνδυασμό με τη δυνατότητα παράλληλης επεξεργασίας ολόκληρης της ακολουθίας, εξηγεί τόσο την ποιότητα των αναπαραστάσεων όσο και την υπολογιστική αποδοτικότητα που επέτρεψε την εκπαίδευση μοντέλων δισεκατομμυρίων παραμέτρων.

Η δημιουργία ενός σύγχρονου γλωσσικού μοντέλου «οδηγιών» δεν είναι, ωστόσο, μονοσταδιακή διαδικασία. Διακρίνονται τρία διαδοχικά στάδια, καθένα από τα οποία προσδίδει διαφορετική ικανότητα. Το πρώτο είναι η προεκπαίδευση (pre-training): το μοντέλο εκτίθεται σε τεράστιους όγκους κειμένου και εκπαιδεύεται να προβλέπει την επόμενη λέξη, μαθαίνοντας με αυτόν τον τρόπο τη στατιστική δομή της γλώσσας, γραμματικά μοτίβα, αλλά και ένα ευρύ φάσμα γνώσεων για τον κόσμο. Το προϊόν αυτού του σταδίου είναι ένα ισχυρό αλλά «ακατέργαστο» μοντέλο, ικανό να συμπληρώνει κείμενο αλλά όχι απαραίτητα να ακολουθεί οδηγίες. Το δεύτερο στάδιο, η προσαρμογή σε οδηγίες (instruction tuning), εκθέτει το μοντέλο σε ζεύγη εντολών και επιθυμητών απαντήσεων, διδάσκοντάς το να ερμηνεύει και να εκτελεί αιτήματα διατυπωμένα σε

φυσική γλώσσα (Wei et al., 2021). Το τρίτο στάδιο, η ευθυγράμμιση (alignment), βελτιώνει περαιτέρω τη συμπεριφορά του μοντέλου ώστε οι απαντήσεις του να είναι χρήσιμες, ασφαλείς και σύμφωνες με τις ανθρώπινες προτιμήσεις· η πλέον διαδεδομένη μέθοδος είναι η ενισχυτική μάθηση από ανθρώπινη ανατροφοδότηση (Ouyang et al., 2022).

Η κατανόηση αυτής της διαδικασίας είναι κρίσιμη για την παρούσα εργασία, καθώς φωτίζει τη φύση του προβλήματος που αντιμετωπίζει το fine-tuning. Ένα μοντέλο «οδηγιών» έχει ήδη μάθει να ακολουθεί γενικές εντολές, δεν έχει όμως εκτεθεί στο ιδιαίτερο πρότυπο του συγκεκριμένου έργου: πώς να συνδυάζει ιστορικές υποδείξεις με την τρέχουσα κατάσταση σε ένα εξειδικευμένο περιβάλλον πλοήγησης. Η προσαρμογή που περιγράφεται στο Κεφάλαιο 4 αποτελεί, υπό αυτή την έννοια, ένα τέταρτο, στοχευμένο στάδιο εξειδίκευσης πάνω σε ένα ήδη ευθυγραμμισμένο μοντέλο.

**Από τα Μοντέλα στους Πράκτορες**

Το επόμενο βήμα στην εξέλιξη των Γλωσσικών Μοντέλων είναι η σύσταση των αυτόνομων πρακτόρων (agents), οι οποίοι στην ουσία τους αντί να παράγουν κείμενο ως τελικό προϊόν, καλούνται να πράξουν ορισμένες ενέργειες όντας σε ένα εξωτερικό περιβάλλον, να παρατηρήσουν στην συνέχεια τα αποτελέσματα των πράξεων αυτών και να προσαρμόσουν την συμπεριφορά τους αναλόγως. Ένας τέτοιος agent με λίγα λόγια δηλαδή, λειτουργεί σε έναν βρόχο αντίληψης – δράσης, όπου λαμβάνοντας μια παρατήρηση (observation) για την τρέχουσα κατάσταση στην οποία βρίσκεται εντός του περιβάλλοντος, αποφασίζει για την επόμενη ενέργειά του, την εκτελεί και εισέρχεται σε μια νέα κατάσταση όπου βλέπει τα αποτελέσματα της δράσης του και καλείται να επαναλάβει αυτήν την διαδικασία.

Στο πλαίσιο αυτό, το γλωσσικό μοντέλο λειτουργεί ως ο μηχανισμός λήψης αποφάσεων (policy) του agent. Η προσέγγιση αυτή είναι ιδιαιτέρως ελκυστική, καθώς αξιοποιεί την ενσωματωμένη γνώση των LLMs, χωρίς να απαιτείται εκ νέου εκπαίδευση για κάθε νέα εργασία. Ωστόσο, έρχεται και με ορισμένες προκλήσεις, καθώς το μοντέλο για να λειτουργεί με αυτονομία αξιόπιστα σε ένα περιβάλλον θα πρέπει να αντιμετωπίσει τις συχνά θορυβώδεις, ημιδομημένες παρατηρήσεις και να διατηρήσει την συνέπεια των αποτελεσμάτων του σε βάθος πολλών βημάτων.

## *2.2 Το πρόβλημα της πλοήγησης στον ιστό και το περιβάλλον WebShop*

### Ορισμός του προβλήματος

Η πλοήγηση στις ιστοσελίδες μπορεί να μοντελοποιηθεί ως μια ακολουθιακή διαδικασία λήψης αποφάσεων όπου σε κάθε χρονική στιγμή t ο agent βρίσκεται σε μια κατάσταση λαμβάνοντας μια παρατήρηση $o_t$ (observation), δηλαδή το περιεχόμενο της τρέχουσας ιστοσελίδας. Στη συνέχεια, δοθέντος ενός στόχου g (goal), διατυπωμένου σε φυσική γλώσσα, ο agent αποφασίζει να πραγματοποιήσει μια ενέργεια $a_t$ (action) την οποία την επιλέγει μέσα από ένα σύνολο διαθέσιμων ενεργειών. Η ενέργεια αυτή εκτελείται και στη συνέχεια ο agent μεταβαίνει σε μια νέα κατάσταση, η οποία είναι απόρροια της ενέργειας που πραγματοποίησε και επιστρέφει τη νέα παρατήρηση $o_{t+1}$. Η διαδικασία αυτή επαναλαμβάνεται έως ότου ολοκληρωθεί ο στόχος, παράγοντας συνεπώς μια τροχιά (trajectory):

$$\tau = (o_1, a_1, o_2, a_2, \dots, o_r, a_r)$$

Στο τέλος του trajectory, το περιβάλλον αποδίδει μια ανταμοιβή (reward) $r \in [0,1]$, η οποία ποσοτικοποιεί το κατά πόσο ο τελικός στόχος g ικανοποιήθηκε επιτυχώς.

### Το περιβάλλον WebShop

Το WebShop σχεδιάστηκε ως ένα ρεαλιστικό αλλά ελεγχόμενο περιβάλλον δοκιμών για πράκτορες που αλληλεπιδρούν με τον ιστό (Yao et al., 2022). Προσομοιώνει ένα ηλεκτρονικό κατάστημα με πάνω από ένα εκατομμύριο πραγματικά προϊόντα, όπου ο agent καλείται να εκπληρώσει μια εντολή αγοράς διατυπωμένη σε φυσική γλώσσα, για παράδειγμα, την εύρεση ενός προϊόντος με συγκεκριμένα χαρακτηριστικά εντός ενός ορίου τιμής. Η επίλυση ενός τέτοιου έργου απαιτεί μια ακολουθία ενεργειών: αναζήτηση με λέξεις-κλειδιά, πλοήγηση στα αποτελέσματα, επιλογή προϊόντος, ρύθμιση χαρακτηριστικών όπως χρώμα ή μέγεθος, και τελική ολοκλήρωση της αγοράς. Ο χώρος ενεργειών περιλαμβάνει, δηλαδή, τόσο ενέργειες αναζήτησης όσο και ενέργειες επιλογής (clicks) πάνω στα διαθέσιμα στοιχεία κάθε σελίδας, ενώ η επιτυχία αποτιμάται με βάση το πόσο η τελική επιλογή ταιριάζει με τις προδιαγραφές της εντολής.

Παράλληλα, το χαρακτηριστικό που καθιστά το WebShop απαιτητικό αλλά και ταυτόχρονα κατάλληλο για τη μελέτη του ρόλου της μνήμης είναι η μερική παρατηρησιμότητά του. Ο agent δεν έχει πρόσβαση στην πλήρη κατάσταση του περιβάλλοντος σε κάθε βήμα· βλέπει μόνο την τρέχουσα σελίδα, χωρίς ρητή αναπαράσταση της διαδρομής που ακολούθησε ή της γνώσης που συγκέντρωσε. Αυτός είναι ο λόγος για τον οποίο το πρόβλημα μοντελοποιείται τυπικά ως μερικώς παρατηρήσιμη μαρκοβιανή διαδικασία απόφασης. Όπως αναλύεται στη συνέχεια, η πλειονότητα των agents αντιμετωπίζει το πρόβλημα με τρόπο ουσιαστικά μαρκοβιανό,

αποφασίζοντας με βάση την τρέχουσα παρατήρηση και μόνο, παραβλέποντας το ιστορικό που, σε ένα μερικώς παρατηρήσιμο περιβάλλον, θα μπορούσε να αποτελεί πολύτιμη πηγή πληροφορίας. Ακριβώς αυτό το κενό αποτελεί το σημείο εκκίνησης της παρούσας εργασίας.

Ακόμα, το WebShop αποτελείται από έναν μεγάλο κατάλογο πραγματικών προϊόντων και συνολικά διαθέτει 5.500 tasks για εκπαίδευση και 500 tasks για αξιολόγηση, με κάθε task να ορίζεται από μια οδηγία σε φυσική γλώσσα που περιγράφει το ζητούμενο προϊόν και τα επιθυμητά χαρακτηριστικά του (χρώμα, μέγεθος, τιμή κτλ.) (Lee et al, 2025).

Ο agent ξεκινώντας από μια αρχική κατάσταση καλείται να προβεί σε μια ακολουθία ενεργειών ώστε να καταλήξει στο σωστό και να ολοκληρώσει την αγορά. Στο τέλος, υπολογίζεται η ανταμοιβή όπως προαναφέρθηκε με βάση το πόσο καλά το τελικό προϊόν ταιριάζει με τις προδιαγραφές του στόχου.

Οι διαθέσιμες ενέργειες στο WebShop ανήκουν σε δύο βασικούς τύπους:

- search[query] → υποβολή ενός ερωτήματος αναζήτησης
- click[element] → επιλογή ενός στοιχείου της σελίδας

**Παράδειγμα observation – action**

Πιο συγκεκριμένα, παρατίθεται ένα ενδεικτικό παράδειγμα. Έστω ο στόχος:

«Ψάχνω για ένα αθλητικό παπούτσι για τρέξιμο σε γκρι χρώμα, με τιμή κάτω από 100 δολάρια.»

Συνεπώς, κατόπιν αναζήτησης και επιλογής προϊόντος, μια τυπική παρατήρηση $o_t$ θα έχει την ακόλουθη μορφή:

**[ Back to Search ]**

**[ < Prev ]**

**color**

**[ black ]**

**[ grey ]**

**Nike Running Shoe….**

**Price: $79,99**

**[ Buy Now ]**

Στην παραπάνω κατάσταση η ορθή ενέργεια που θα έπρεπε να κάνει ο agent θα ήταν η επιλογή του σωστού χρώματος συνεπώς:

Click[grey]

## *2.3 Fine-tuning και τεχνική LoRA*

### Η Έννοια του fine-tuning

Όπως περιεγράφηκε και προηγουμένως, ένα LLM αποκτά τις ικανότητές του κατά το στάδιο της προεκπαίδευσης, στο οποίο εκτίθεται σε τεράστιες ποσότητες γενικού κειμένου. Ωστόσο, αυτό είναι λιγότερο αποτελεσματικό όταν θέλουμε να το χρησιμοποιήσουμε για μια εξειδικευμένη εργασία, καθώς οι πληροφορίες στις οποίες εκτίθεται είναι γενικής φύσεως και ως εκ τούτου δεν είναι κατάλληλα εκπαιδευμένο για συγκεκριμένες εργασίες. Για να αντιμετωπιστεί αυτό το πρόβλημα, οι ερευνητές ανακάλυψαν μια διαδικασία που την ονόμασαν fine-tuning, κατά την οποία το ήδη προεκπαιδευμένο μοντέλο εκπαιδεύεται περαιτέρω, για ένα επιπλέον διάστημα πάνω σε ένα μικρότερο και στοχευμένο σύνολο δεδομένων με αποτέλεσμα το μοντέλο να προσαρμοστεί στις ιδιαιτερότητες της εκάστοτε εφαρμογής, χωρίς να εκπαιδευτεί από την αρχή.

Στο πλαίσιο της παρούσας εργασίας, το fine-tuning είναι το εργαλείο με το οποίο διδάσκουμε στο μοντέλο απόφασης πώς να αξιοποιεί σωστά τα hints που του παρέχει το GST.

### Το πρόβλημα του πλήρους fine-tuning

Παραδοσιακά, το fine-tuning αφορούσε την ενημέρωση όλων των παραμέτρων του μοντέλου κατά τη διάρκεια της εκπαίδευσης (full fine-tuning), ωστόσο κάτι τέτοιο όταν πρόκειται για μοντέλα με δισεκατομμύρια παραμέτρους γίνεται ιδιαίτερα κοστοβόρο και χρονοβόρο, καθώς απαιτεί μεγάλη ποσότητα μνήμης GPU για να αποθηκευτούν και ενημερωθούν οι παράμετροι , αλλά και διαρκούσε αρκετές ώρες εώς και ημέρες. Συνεπώς, για πολλούς ερευνητές, ιδίως όταν οι πόροι είναι περιορισμένοι, το πλήρες fine-tuning είναι πρακτικά ανέφικτο.

**LoRA: μια αποδοτική προσέγγιση**

Για να αντιμετωπιστεί ο παραπάνω περιορισμός, αναπτύχθηκε μια οικογένεια τεχνικών αποδοτικών ως προς τις παραμέτρους προσαρμογής (Parameter-Efficient Fine-Tuning, PEFT), με την πιο διαδεδομένη από αυτές να ονομάζεται LoRA (Low-Rank Adaptation). (Hu et al., 2022).

Συνεπώς, αντί να χρησιμοποιούνται όλοι οι παράμετροι για την εκπαίδευση του μοντέλου, αυτές «κλειδώνουν» (παραμένουν αμετάβλητες) και η προσαρμογή επιτυγχάνεται με την προσθήκη ενός μικρού αριθμού νέων παραμέτρων σε επιλεγμένα σημεία της αρχιτεκτονικής. Οι παράμετροι αυτοί είναι ουσιαστικά δύο μικροί πίνακες χαμηλής τάξης (low-rank), οι οποίες όταν πολλαπλασιαστούν, προσεγγίζουν την αναγκαία μεταβολή στη συμπεριφορά του μοντέλου. Οπότε κατά το στάδιο της εκπαίδευσης ενημερώνονται μόνο αυτοί οι μικροί πίνακες, ενώ το υπόλοιπο μοντέλο παραμένει αμετάβλητο.

Το αποτέλεσμα της παραπάνω διαδικασίας είναι να μειωθεί σημαντικά ο αριθμός των εκπαιδεύσιμων παραμέτρων και κατά συνέπεια να επιτυγχάνονται μικρότερες απαιτήσεις μνήμης, ταχύτερη εκπαίδευση και ελάχιστο αποθηκευτικό κόστος.

Η αποτελεσματικότητα, δηλαδή, της LoRA στηρίζεται στο γεγονός πως η μεταβολή που απαιτείται για την εξειδίκευση ενός ήδη εκπαιδευμένου μοντέλου σε ένα νέο έργο τείνει να ζει σε έναν χώρο πολύ μικρότερης διάστασης από το σύνολο των παραμέτρων του μοντέλου. Με άλλα λόγια, δεν χρειάζεται να μεταβληθεί ουσιαστικά κάθε παράμετρος· η αναγκαία προσαρμογή μπορεί να προσεγγιστεί ικανοποιητικά από έναν μικρό αριθμό νέων παραμέτρων. Αυτή η υπόθεση εξηγεί γιατί η ενημέρωση ενός ελάχιστου κλάσματος των παραμέτρων επαρκεί για να αποδώσει το μοντέλο σε εξειδικευμένες εργασίες, χωρίς αισθητή απώλεια σε σχέση με την πλήρη επανεκπαίδευση.

Πέραν της υπολογιστικής οικονομίας, η προσέγγιση αυτή προσφέρει και πρακτικά πλεονεκτήματα ιδιαίτερα χρήσιμα για την παρούσα εργασία, διότι καθώς το βασικό μοντέλο παραμένει αμετάβλητο και η προσαρμογή αποθηκεύεται ως ένα ξεχωριστό, μικρό σύνολο παραμέτρων, καθίσταται δυνατή η εναλλαγή μεταξύ του βασικού και του fine-tuned μοντέλου με ελάχιστο κόστος, μια δυνατότητα που αξιοποιήθηκε άμεσα στην πειραματική αξιολόγηση, όπου οι δύο εκδοχές του μοντέλου συγκρίνονται υπό ταυτόσημες συνθήκες. Επιπλέον, το μικρό μέγεθος της προσαρμογής διευκολύνει την αποθήκευση, τη μεταφορά και την αναπαραγωγή του πειράματος, ενισχύοντας τη συνολική μεθοδολογική διαφάνεια.

Χάρη στα παραπάνω πλεονεκτήματα, η τεχνική LoRA επιλέχθηκε για την εκπαίδευση του μοντέλου απόφασης στην παρούσα εργασία, καθιστώντας εφικτή την προσαρμογή ενός μοντέλου δισεκατομμυρίων παραμέτρων.

## *2.4 Η μέθοδος LCoW*

Η μέθοδος LCoW (Lee et al., 2025) αποτελεί το άμεσο σημείο αναφοράς της παρούσας εργασίας. Η μέθοδος αυτή αναπτύχθηκε για να αντιμετωπίσει μια σημαντική δυσκολία όσον αφορά την αυτόνομη πλοήγηση των agents σε ένα περιβάλλον ιστού, η οποία αφορά το γεγονός πως το περιεχόμενο μιας ιστοσελίδας είναι συχνά αρκετά θορυβώδες, εκτενές και συνεπώς καθίσταται δύσκολο να ερμηνευτεί σωστά και να ληφθούν οι σωστές ενέργειες από τους agents. Ως εκ τούτου, η μέθοδος αυτή προτείνει ένα ενδιάμεσο στάδιο επεξεργασίας του περιεχομένου ώστε να δοθεί μια καθαρότερη, πιο περιεκτική έκδοσή του στο μοντέλο πριν την λήψη της απόφασης.

### To Contextualization Module

Στην καρδιά του LCoW βρίσκεται ένα ξεχωριστό νευρωνικό δίκτυο, το module εννοιολογικής πλαισίωσης όπως το αναφέρουν (contextualization module). Ο ρόλος του είναι να λαμβάνει το ακατέργαστο observation της σελίδας και να το αναδιατυπώνει σε μια πιο καθαρή, συνεκτική μορφή, κρατώντας μόνο τις πληροφορίες που είναι σχετικές με τον στόχο που έχει δοθεί. Στη συνέχεια, αυτό το νέο observation δίνεται στο LLM, το οποίο λαμβάνει την τελική απόφαση και συνεπώς βελτιώνεται κατά πολύ η ακρίβεια των επιλογών του δεδομένου ότι του έχει δοθεί μια πιο ευανάγνωστη και περιεκτική έκδοση του observation ως είσοδο.

### Reward based contextualization

Δεδομένου της απουσίας έτοιμων, πρότυπων αναδιατυπώσεων των observations, το LCoW εισάγει μια ενδιαφέρουσα μέθοδο για την εκπαίδευση του module ώστε να παράγει χρήσιμες αναδιατυπώσεις. Αυτή αφορά τον ορισμό της ποιότητας της αναδιατύπωσης, όχι με βάση το πώς φαίνεται οπτικά, αλλά με βάση το πόσο χρήσιμη αποδεικνύεται στην πράξη για το μοντέλο. Πιο συγκεκριμένα, μια αναδιατύπωση θεωρείται καλή αν, όταν δοθεί στο LLM, το οδηγεί να επιλέξει την σωστή ενέργεια.

Στα πλαίσια αυτής της λογικής, η διαδικασία εκπαίδευσης ακολουθεί μια προσέγγιση καθοδηγούμενο από τον ορισμό μιας ανταμοιβής (reward-based approach) αναλόγως αν η αναδιατύπωση είναι καλή ή λιγότερο καλή. Συνεπώς, για κάθε κατάσταση, το module παράγει πολλές υποψήφιες αναδιατυπώσεις οι οποίες αξιολογούνται ως προς το αν

οδηγούν στην σωστή ενέργεια και όσες επιτυγχάνουν χρησιμοποιούνται ως θετικά παραδείγματα για την περαιτέρω εκπαίδευση του module. Ως αποτέλεσμα αυτού, το module μαθαίνει σταδιακά να παράγει αναδιατυπώσεις που μεγιστοποιούν την πιθανότητα σωστής ενέργειας.

**Επαναληπτική βελτίωση**

Η προαναφερθείσα διαδικασία επαναλαμβάνεται σε διαδοχικούς κύκλους (iterations), όπου σε κάθε κύκλο, το βελτιωμένο module χρησιμοποιείται για να παραχθούν οι νέες υποψήφιες αναδιατυπώσεις που με την σειρά τους τροφοδοτούν τον επόμενο γύρο εκπαίδευσης. Αυτή η επαναληπτική λογική επιτρέπει στο σύστημα να βελτιώνεται με την πάροδο των επαναλήψεων αφού κάθε κύκλος βασίζεται στις ικανότητες του προηγούμενου, οδηγώντας σε ολοένα και καλύτερα contextualized observations.

**Η σχέση με την παρούσα εργασία**

Η παρούσα εργασία έχει εμπνευστεί άμεσα από την μεθοδολογία της LCoW και ιδίως από την λογική της καθοδηγούμενης από ανταμοιβή επιλογής , ωστόσο , όπως θα αναλυθεί και στο Κεφάλαιο 4, διαφοροποιείται ουσιωδώς ως προς το πού εφαρμόζεται αυτή η προσέγγιση : αντί να εκπαιδεύεται ένα νευρωνικό module που αναδιατυπώνει την είσοδο, η προσέγγισή μας αξιοποιεί μια συμβολική δομή, αυτή του Γενικευμένου Δέντρου Επιθημάτων, για να καθοδηγήσει απευθείας το στάδιο της απόφασης.

## *2.5 Γενικευμένο Δέντρο Επιθημάτων (Generalized Suffix Tree, GST)*

**Επιθήματα και Δέντρα Επιθημάτων**

Δεδομένου να γίνει κατανοητή η δομή που βρίσκεται στον πυρήνα της προσέγγισής μας, θα πρέπει να γίνει κατανοητή μια θεμελιώδη έννοια, αυτή των επιθημάτων. Δοθείσης μιας ακολουθίας συμβόλων , συνεπώς, ορίζουμε ως επίθημα (suffix) οποιοδήποτε τμήμα που εκτείνεται από κάποια θέση μέχρι το τέλος της ακολουθίας αυτής. Παραδείγματος χάριν, η ακολουθία «ABC» έχει τα εξής επιθήματα: «ABC», «BC», «C».

Ένα Δέντρο Επιθημάτων (Suffix Tree) (Weiner, 1973), λοιπόν, είναι μια δενδρική δομή δεδομένων όπου αποθηκεύει αυτά τα επιθήματα μιας ακολουθίας με οργανωμένο και αποδοτικό τρόπο. Συνιστούν, δηλαδή, μια από τις θεμελιώδεις δομές δεδομένων της επεξεργασίας συμβολοσειρών, με ιστορία που εκτείνεται σε πέντε δεκαετίες. Η πρώτη γραμμική κατασκευή προτάθηκε από τον Weiner (1973), ενώ αργότερα ο Ukkonen (1995) εισήγαγε έναν διαδικτυακό (online) αλγόριθμο κατασκευής που παραμένει ο πλέον διαδεδομένος. Η κεντρική τους ιδιότητα είναι ότι αναπαριστούν όλα τα επιθήματα μιας συμβολοσειράς σε μια συμπαγή δενδρική δομή, επιτρέποντας τον εντοπισμό οποιουδήποτε υπο-μοτίβου σε χρόνο ανάλογο του μήκους του μοτίβου και όχι του μήκους ολόκληρης της συμβολοσειράς. Η ιδιότητα αυτή τα καθιστά ιδανικά για εφαρμογές όπου απαιτείται επαναλαμβανόμενη αναζήτηση προτύπων, από τη βιοπληροφορική έως τη συμπίεση δεδομένων.

Η βασική λογική πίσω από τη δομή είναι η εξής: αντί να αποθηκεύεται κάθε επίθημα χωριστά, κάτι που θα απαιτούσε τετραγωνικό χώρο, τα κοινά προθέματα των επιθημάτων μοιράζονται τους ίδιους κόμβους, με αποτέλεσμα η δομή να συμπτύσσει αυτόματα κάθε επαναλαμβανόμενο τμήμα. Ο αλγόριθμος του Ukkonen επιτυγχάνει την κατασκευή σε αποσβεσμένο γραμμικό χρόνο μέσω τριών βασικών τεχνικών: της επεξεργασίας της ακολουθίας σύμβολο προς σύμβολο διατηρώντας ένα «ενεργό σημείο» που αποφεύγει την επανάληψη διαδρομών, της χρήσης «συνδέσμων επιθημάτων» που επιτρέπουν ταχεία μετάβαση μεταξύ σχετικών κόμβων, και της άμεσης επέκτασης των φύλλων χωρίς επιπλέον κόστος.

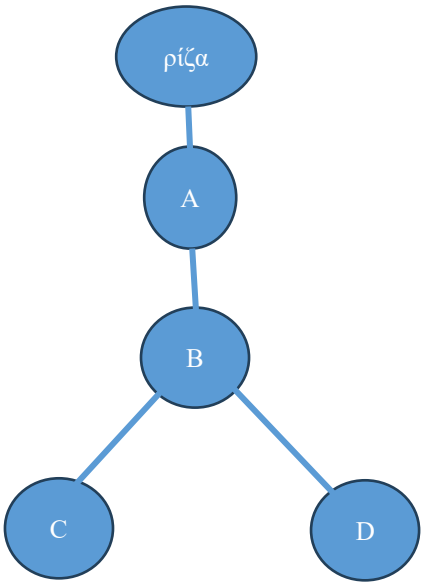
Η γενίκευση επιτυγχάνεται με την προσθήκη ενός μοναδικού διαχωριστικού συμβόλου στο τέλος κάθε ακολουθίας, ώστε τα επιθήματα διαφορετικών ακολουθιών να μη συγχέονται μεταξύ τους. Στο πλαίσιο της παρούσας εργασίας, αυτή ακριβώς η ιδιότητα είναι που επιτρέπει την ενοποίηση όλων των ιστορικών trajectories σε μία δομή: κάθε επαναλαμβανόμενο μοτίβο καταστάσεων, ανεξαρτήτως του trajectory στο οποίο εμφανίστηκε ή της θέσης του μέσα σε αυτό, εντοπίζεται και ομαδοποιείται αυτόματα, καθιστώντας δυνατή την αποδοτική ανάκτηση ιστορικών προηγουμένων κατά τη λήψη απόφασης. Στο δέντρο αυτό, κάθε διαδρομή ξεκινώντας από την ρίζα, ακολουθώντας

τους κόμβους προς τα φύλλα του δέντρου αποτελεί ένα κομμάτι της αρχικής ακολουθίας. Επίσης, όταν δύο επιθήματα ξεκινούν με τα ίδια σύμβολα τότε δεν αποθηκεύονται δύο φορές, αλλά μοιράζονται την ίδια αρχική διαδρομή στο δέντρο και αυτό διακλαδώνεται στο σημείο που αρχίζουν να διαφέρουν.

Το κέρδος σε αυτή την οργάνωση είναι η ταχύτητα, καθώς αντί να χρειάζεται να διατρέξουμε ολόκληρη την ακολουθία κάθε φορά που θέλουμε να ελέγξουμε αν υπάρχει ένα συγκεκριμένο μοτίβο, τότε το δέντρο μας επιτρέπει να το εντοπίσουμε γρηγορότερα, ακολουθώντας απλώς την αντίστοιχη διαδρομή.

**Ενδεικτικό Παράδειγμα**

Έστω ένα σύνολο δύο ακολουθιών: «ABC» και «ABD». Παρατηρούμε πως και οι δύο ξεκινούν με το τμήμα «AB» και διαφοροποιούνται στο τελευταίο, συνεπώς σε ένα GST το κοινό τμήμα «AB» θα αναπαρασταθεί μία φορά ως κοινή διαδρομή που ξεκινά από την ρίζα και θα υπάρχουν στη συνέχεια δύο κλάδοι ένας που καταλήγει στο «C» και ένας στο «D».




**Εικόνα 2.1:** Παράδειγμα δομής ενός Γενικευμένου Δέντρου Επιθημάτων για τις ακολουθίες «ABC» και «ABD»

Σημαντική λεπτομέρεια που αξίζει να σημειωθεί είναι πως σε κάθε κόμβο δίνεται η δυνατότητα να μπορεί να διατηρηθεί η πληροφορία του πόσες φορές διασχίστηκε η αντίστοιχη διαδρομή, προσδίδοντας ένα μέτρο συχνότητας και συνεπώς αξιοπιστίας κάθε μοτίβου, γεγονός ιδιαίτερα χρήσιμο για την προσέγγισή μας.

**Επιλογή του Γενικευμένου Δέντρου Επιθημάτων για την προσέγγισή μας**

Η καταλληλότητα του GST για την προσέγγισή μας προκύπτει άμεσα από τα προαναφερθέν χαρακτηριστικά του. Στην περίπτωση της αυτόνομης πλοήγησης των agents στον Ιστό, κάθε trajectory μπορεί να κωδικοποιηθεί ως μια ακολουθία, μια αλληλουχία δηλαδή καταστάσεων και ενεργειών. Συνεπώς, αν συγκεντρώσουμε όλες αυτές τις διαδρομές εντός ενός GST θα έχουμε καταφέρει να αποκτήσουμε μια συμπαγή οργανωμένη αναπαράσταση όλων των ιστορικών μοτίβων.

Με αυτόν τον τρόπο, δηλαδή, δίνεται απάντηση στο ερώτημα του ποιες ενέργειες ακολουθήθηκαν στο παρελθόν σε παρόμοια περίσταση και πόσο συχνά. Επιπλέον, χάρη στην ιεραρχική δομή του δέντρου, επιτρέπεται η σταδιακή αναζήτηση ενός μοτίβου ούτως ώστε αν ένα συγκεκριμένο μοτίβο δεν έχει παρατηρηθεί στο παρελθόν, η δομή αυτή μπορεί να «σπάσει» σε μικρότερα μοτίβα που είναι συχνότερα, μια ιδιαιτερότητα ιδιαίτερα καθοριστική στην προσέγγισή μας όπου θα αναλυθεί εκτενώς στο Κεφάλαιο 4. Συνεπώς, το πλεονέκτημα αυτής της δομής έναντι ενός νευρωνικού δικτύου έγγυται στο ότι η γνώση δεν είναι κρυμμένη μέσα σε παραμέτρους, αντ' αυτού σε κάθε μοτίβο η συχνότητα του είναι ρητά και άμεσα διαθέσιμη, καθιστώντας την δομή αυτή πλήρως ερμηνεύσιμη.

# 3

# Σχετική Βιβλιογραφία

## *3.1 Πράκτορες πλοήγησης Ιστού*

Τα τελευταία χρόνια με την ραγδαία πρόοδο που υπάρχει στον τομέα της Τεχνητής Νοημοσύνης, ιδιαίτερο ερευνητικό ενδιαφέρον έχει παρατηρηθεί στο πρόβλημα της αυτόνομης πλοήγησης των agents σε ιστοσελίδες έχοντας ως αποτέλεσμα την ανάπτυξη πλήθους περιβαλλόντων αξιολόγησης (benchmarks) για την μέτρηση των ικανοτήτων τους. Οι πρώτες σχετικές προσπάθειες όπως το MiniWoB (Shi et al., 2017) και η επέκτασή του MiniWoB++ (Liu et al., 2018) εισήγαγαν ορισμένα ελεγχόμενα benchmarks για την αξιολόγηση βασικών αλληλεπιδράσεων όπως είναι η συμπλήρωση φορμών και το πάτημα κουμπιών από τους agents.

Ενώ τα benchmarks αυτά καθόρισαν και άνοιξαν τον δρόμο για περαιτέρω έρευνες που εξέλιξαν τους agents, περιορίζονταν σε σχετικά απλές εργασίες. Ως αποτέλεσμα αυτού, δημιουργήθηκε η ανάγκη για πιο σύνθετα και ρεαλιστικά benchmarks όπως ένα από αυτά είναι το WebShop (Yao et al., 2022), το οποίο αποτελεί και το περιβάλλον αξιολόγησης της παρούσας εργασίας. Το WebShop προσομοιώνει ένα ηλεκτρονικό κατάστημα με πραγματικά δεδομένα προϊόντων, όπου ο agent καλείται να εκτελέσει ορισμένες ενέργειες και να καταλήξει σε αγορά ενός προϊόντος βάσει οδηγιών δοσμένων σε φυσική γλώσσα. Παράλληλα, άλλες προσεγγίσεις επέκτειναν το πεδίο προς μεγαλύτερη ποικιλία και ρεασλισμό: το Mind2Web (Deng et al., 2023) συγκέντρωσε εργασίες από πλήθος πραγματικών ιστοσελίδων, ενώ το WebArena (Zhou et al., 2024) εισήγαγε ένα πλήρως λειτουργικό περιβάλλον με πολλαπλούς τομείς εφαρμογής, δίνοντας έμφαση στη λειτουργική ορθότητα των ενεργειών του agent.

Αξίζει, ακόμα, να σημειωθεί πως σε πολλά benchmarks όπως είναι και το WebShop έχει παρατηρηθεί πως η μοντελοποίηση της πλοήγησης του agent αποτελεί μια μερικώς παρατηρήσιμη μαρκοβιανή διαδικασία απόφασης (Partially-Observable Markov Decision Process, POMDP), σημαίνοντας πως η επόμενη ενέργεια εξαρτάται αποκλειστικά από την τρέχουσα κατάσταση, χωρίς ρητή αξιοποίηση του ιστορικού των

προηγούμενων αλληλεπιδράσεων. Η απουσία αυτής της ρητής μνήμης του ιστορικού είναι που η παρούσα εργασία επιχειρεί να συνεισφέρει και να επιλύσει.

## *3.2 Επεξεργασία των observations και Contextualization Module*

Όπως έχει αναφερθεί και προηγουμένως, ένα από τα πιο δύσκολα προβλήματα που καλούνται να επιλύσουν οι ερευνητές όσον αφορά την αυτόνομη πλοήγηση των agents στον Ιστό είναι το γεγονός πως το ακατέργαστο περιεχόμενο μιας ιστοσελίδας , είτε πρόκειται για κώδικα HTML είτε για δέντρο προσβασιμότητας (accessibility tree) , είναι κατά κανόνα υπερβολικά εκτενές και θορυβώδες. Ως αποτέλεσμα αυτού εγείρονται δύο προβλήματα: αφενός όταν δίνεται ως είσοδο σε γλωσσικά μοντέλα τότε αυτή προφανώς θα υπερβαίνει τα όρια εισόδου και αφετέρου η αποτελεσματικότητα όσον αφορά την λήψη σωστών αποφάσεων μειώνεται δραστικά, καθώς η σχετική πληροφορία χάνεται μέσα σε πλήθος άσχετων στοιχείων. Για τον λόγο αυτό, ένα σημαντικό μέρος της βιβλιογραφίας έχει επικεντρωθεί σε τεχνικές που μειώνουν ή αναδιατυπώνουν το αρχικό observation πριν αυτό δοθεί στο μοντέλο που θα λάβει την απόφαση.

Οι προσεγγίσεις αυτές μπορούν να κατηγοριοποιηθούν σε ορισμένες κατηγορίες όπως περιγράφεται ακολούθως:

Μια πρώτη οικογένεια μεθόδων στηρίζεται στην επιλογή ή φιλτράρισμα των στοιχείων μιας σελίδας, π.χ. οι Deng et al. (2023) εκπαιδεύουν έναν βαθμολογητή που εκτιμά τη συνάφεια κάθε στοιχείου της σελίδας με βάση την οδηγία που έχει δοθεί και στη συνέχεια προωθεί στο μοντέλο μόνο τα πιο σχετικά. Παρόμοιας λογικής είναι και μέθοδοι που αφαιρούν περιττά σημεία του κώδικα μιας σελίδας (DOM pruning), διατηρώντας με αυτόν τον τρόπο μια πιο συμπαγή κατάσταση. Πέραν αυτών, μια δεύτερη οικογένεια μεθόδων αξιοποιεί την σύνοψη της κατάστασης όπως π.χ. οι Gur et al.(2024) όπου εισάγουν μια διαδικασία δύο σταδίων, όπου ένα εξειδικευμένο μοντέλο εξάγει και κάνει ουσιαστικά μια περίληψη των σχετικών στοιχείων πριν αυτά περάσουν στο μοντέλο.

Στο ίδιο πλαίσιο με το τελευταίο εντάσσεται και η μέθοδος LCoW (Lee et al., 2025), η οποία αποτελεί το άμεσο σημείο αναφοράς της παρούσας εργασίας, όπου αντί απλώς να φιλτράρει το περιεχόμενο μιας ιστοσελίδας, η LCoW εκπαιδεύει ένα ξεχωριστό μοντέλο που αναδιατυπώνει το observation σε μια πιο κατανοητή μορφή, περιγράφοντας σε φυσική γλώσσα τον ρόλο των στοιχείων της σελίδας. Παρατηρήθηκε, συνεπώς πως κοινό χαρακτηριστικό αυτών των προσεγγίσεων είναι πως επικεντρώνονται στο

προπαρασκευαστικό στάδιο, στο πως δηλαδή θα βελτιωθεί η ποιότητα της εισόδου που φτάνει στο μοντέλο απόφασης.

Η παρούσα εργασία, συνεπώς, διαφοροποιείται σε αυτό ακριβώς το σημείο. Αντί να επενδύσουμε στη βελτίωση της παρατήρησης μέσω ενός επιπλέον νευρωνικού μοντέλου , που προσθέτει πολυπλοκότητα και κόστος, στρέφουμε την προσοχή μας στο ίδιο το στάδιο της απόφασης, εισάγοντας μια συμβολική δομή δεδομένων η οποία καθοδηγεί ουσιαστικά το μοντέλο απόφασης αξιοποιώντας το ιστορικό των προηγούμενων πλοηγήσεων , μια διάσταση που απουσιάζει από τις προαναφερθείσες μεθόδους

## *3.3 Συμβολικές και Υβριδικές Προσεγγίσεις*

Όπως αναφέρθηκε και προηγουμένως, το περιβάλλον WebShop μπορεί να μοντελοποιηθεί ως μια παρατηρήσιμη διαδικασία απόφασης όπου οι περισσότεροι agents που δρουν στο περιβάλλον αυτό λειτουργούν Μαρκοβιανά, δηλαδή αποφασίζουν την επόμενη ενέργειά τους με βάση την τρέχουσα παρατήρηση, χωρίς ρητή αναπαράσταση του ιστορικού των προηγούμενων πλοηγήσεων. Συνεπώς, σύγχρονες μέθοδοι που βασίζονται σε agents με εξωτερική μνήμη προηγούμενων καταστάσεων την οποία ανακαλούν κατά την τρέχουσα λήψη απόφασης, προσπαθούν να καλύψουν αυτό το κενό στην Μαρκοβιανή διαδικασία.

Μια από αυτές τις μεθόδους που αποθηκεύουν τις καταστάσεις ως ελεύθερο κείμενο σε φυσική γλώσσα είναι και το Reflexion (Shinn et al., 2023), το οποίο διατηρεί λεκτική ανατροφοδότηση από τις αποτυχημένες προσπάθειες ώστε να βελτιώσει τις μελλοντικές. Ακόμα, το ExpeL (Zhao et al., 2024) είναι μια αντίστοιχη μέθοδος που εξάγει επαναχρησιμοποιήσιμα αποτελέσματα από επιτυχημένα και αποτυχημένα trajectories. Άλλες προσεγγίσεις ανακτούν σχετικές καταστάσεις από το παρελθόν τη στιγμή της απόφασης όπως π.χ. το Retrieval-Augmented Planning όπου ανακτά observations που αντιστοιχούν στην τρέχουσα κατάσταση για να ενισχύσει την διαδικασία λήψης απόφασης (Kagaya et al., 2024), ενώ υπάρχουν και Generative Agents όπου ανακαλούν αναμνήσεις ταξινομημένες από τις πιο πρόσφατες προς τις παλαιότερες, αλλά και με βάση τη σημασία και τη συνάφειά τους (Park et al., 2023). Κοινό γνώρισμα όλων αυτών είναι πως η μνήμη παραμένει μη δομημένη και χρησιμοποιείται το LLM ως διαμεσολαβητής.

Περεταίρω έρευνα ανέδειξε και μια ακόμα κατηγορία μεθόδων όπου κινείται προς πιο δομημένες αναπαραστάσεις. Συνεπώς, το Agent Workflow Memory χρησιμοποιεί επαναλαμβανόμενες ρουτίνες (workflows) από παρελθοντικά trajectories και τις παρέχει επιλεκτικά στον agent ως καθοδήγηση (Wang et al., 2024). Ακόμα, το Synapse εισάγει

μια αφαίρεση κατάστασης, φιλτράροντας πληροφορία άσχετη με το εκάστοτε task, αλλά και χρησιμοποιεί πλήρη trajectories ως παραδείγματα για προτροπή στην ενέργεια (Zheng et al., 2024). Οι προσεγγίσεις αυτές δείχνουν πως η επαναχρησιμοποίηση της «εμπειρίας» βελτιώνει σημαντικά τη συμπεριφορά, ωστόσο η μνήμη ακόμα παραμένει ελεύθερης μορφής, γεγονός που καθιστά τη συντήρησή της καθώς μεγαλώνει αρκετά δύσκολη, συνυπάρχοντας με ζητήματα παλαίωσης, πλεονασμού και εσωτερικής ασυνέπειας.

Παράλληλα, το ευρύτερο πεδίο των υβριδικών προσεγγίσεων επιχειρεί να συνδυάσει τη στατιστική ικανότητα αναγνώρισης προτύπων των νευρωνικών δικτύων με τη διαφάνεια και τον ελεγχόμενο χαρακτήρα των συμβολικών αυτών δομών (Fang et al., 2024· Zheng et al., 2022 — JARVIS). Στα πλαίσια της λογικής αυτής, μια συμβολική δομή δεν αντικαθιστά το LLM αλλά το πλαισιώνει, προσφέροντας ντετερμινιστική πληροφορία που μπορεί να αναπαραχθεί άμεσα και εύκολα.

Η παρούσα εργασία τοποθετείται σε αυτό ακριβώς το σημείο τομής, όπου αντί για μνήμη που έχει εκπαιδευτεί, χρησιμοποιεί ένα Γενικευμένο Δέντρο Επιθημάτων ως ρητή συμβολική δομή που κωδικοποιεί ακολουθίες observation IDs από trajectories. Επιπλέον, σε αντίθεση με τις παραπάνω προσεγγίσεις που παρεμβαίνουν είτε στην ανάκτηση της εμπειρίας του agent είτε στο προπαρασκευαστικό στάδιο του contextualization των παρατηρήσεων, η δομή αυτή παρεμβαίνει στο στάδιο της απόφασης, παρέχοντας ιστορικές υποδείξεις ενώ το γλωσσικό μοντέλο διατηρεί πάντα τον έλεγχο της τελικής απόφασης.

## *3.4 Θέση της παρούσας Εργασίας*

Συνοψίζοντας τις προηγούμενες ενότητες, σύμφωνα με τις πρόσφατες μελέτες όσον αφορά την αυτόνομη πλοήγηση των agents στον Ιστό, έχουν αναδειχθεί δύο κυρίαρχες κατευθύνσεις στην ενίσχυση της πλοήγησής τους. Η πρώτη επεμβαίνει στην επεξεργασία των observations φιλτράροντας , συνοψίζοντας ή αναδιατυπώνοντας την είσοδο πριν αυτή φτάσει στο μοντέλο απόφασης. Η δεύτερη χρησιμοποιεί μια μνήμη που κατέχει τις παρελθοντικές ενέργειες του agent , την οποία και ανακαλεί για να καθοδηγήσει τις τρέχουσες ενέργειες. Οι δύο αυτές κατευθύνσεις, ωστόσο, αφήνουν ακάλυπτο ένα κενό το οποίο αφορά πως οι agents παραμένουν κατά βάση Μαρκοβιανοί, χωρίς ρητή, δομημένη αναπαράσταση του ιστορικού που να επενεργεί τη στιγμή της απόφασης.

Αυτό είναι και το σημείο διαφοροποίησης της παρούσας εργασίας , η οποία διαφοροποιείται από την βιβλιογραφία κατά τρεις άξονες: Ως προς το στάδιο της παρέμβασης όπου δεν επενεργεί στο προπαρασκευαστικό στάδιο , όπως οι προσεγγίσεις επεξεργασίας των observations και ειδικότερα το LCoW που αναδιατυπώνει τις παρατηρήσεις (Lee et al., 2025),αλλά απευθείας στο στάδιο της απόφασης, όπου η ιστορική πληροφορία παρέχεται ως υπόδειξη στο μοντέλο με στόχο την καθοδήγησή του. Ακόμα, ως προς την μορφή της μνήμης δεν χρησιμοποιεί αναπαραστάσεις ελεύθερου κειμένου, όπως η λεκτική ανατροφοδότηση, τα εξαγόμενα αποτελέσματα ή τα επαγόμενα workflows (Shinn et al., 2023 · Zhao et al., 2024 · Wang et al., 2024), αλλά χρησιμοποιεί μια ρητή ντετερμινιστική συμβολική δομή, το Γενικευμένο Δέντρο Επιθημάτων που κωδικοποιεί ακολουθίες Observation IDs με τρόπο που είναι άμεσα αναπαραγώγιμος. Τέλος, ως προς τον ρόλο του γλωσσικού μοντέλου, η συμβολική δομή δεν το αντικαθιστά ούτε το παρακάμπτει, αντ' αυτού το μοντέλο διατηρεί πάντα τον λόγο της τελικής απόφασης ενώ η δομή έχει έναν συμβουλευτικό ρόλο.

Η σχέση με το LCoW αξίζει ιδιαίτερης ρητής διατύπωσης, καθώς αποτελεί την άμεση αφετηρία της εργασίας , καθώς οι δύο προσεγγίσεις μοιράζονται το ίδιο υποκείμενο πρόβλημα που είναι η ενίσχυση ενός agent ως προς την λήψη της σωστής ενέργειας σε ένα περιβάλλον, που επεμβαίνουν ωστόσο σε διαφορετικά σημεία της αλυσίδας. Το μεν LCoW βελτιώνει το τι αναπαραστάσεις παρατηρεί το μοντέλο, ενώ η παρούσα εργασία επάγεται στο να ενισχύσει το μοντέλο ως προς μια καλύτερη λήψη απόφασης παρέχοντας το ιστορικό ως καθοδηγητή. Υπό αυτήν την έννοια, οι δύο αυτές παρεμβάσεις είναι συμπληρωματικές παρά ανταγωνιστικές.

Η αρχιτεκτονική που υλοποιεί την υβριδική αυτή προσέγγιση, δηλαδή η δομή του GST, ο μηχανισμός παραγωγής υποδείξεων , η προσαρμογή του μοντέλου μέσω fine-tuning και η δρομολόγηση της απόφασης αναπτύσσονται αναλυτικά στο Κεφάλαιο 4.

# 4

# Αρχιτεκτονική

## 4.1 Επισκόπηση Αρχιτεκτονικής

Στο Κεφάλαιο αυτό θα αναλυθεί η αρχιτεκτονική του υβριδικού συστήματος που προτείνουμε. Προτού ξεκινήσουμε την αναλυτική περιγραφή, αξίζει να διατυπωθεί το πρόβλημα που καλείται να λύσει η προσέγγιση αυτή, όπως προκύπτει και από το Κεφάλαιο 3, όπου οι agents πλοήγησης στο WebShop σύμφωνα με το LCoW λειτουργούν κατά κανόνα Μαρκοβιανά, γεγονός που σημαίνει πως δεν έχουν πρόσβαση σε κάποια δομημένη γνώση για το πώς εξελίχθηκαν παρόμοιες καταστάσεις σε προηγούμενα trajectories. Η προτεινόμενη αρχιτεκτονική, συνεπώς, καλύπτει το κενό αυτό εισάγοντας την συμβολική δομή του Γενικευμένου Δέντρου Επιθημάτων, ούτως ώστε να αναπαρασταθεί πλήρως η ιστορικότητα στο στάδιο της απόφασης.

Ο υβριδικός χαρακτήρας της προσέγγισης αυτής λοιπόν, προκύπτει από το γεγονός πως συνδυάζονται δύο ετερογενή συστατικά με διακριτούς ρόλους το καθένα. Το πρώτο είναι το Γενικευμένο Δέντρο Επιθημάτων (GST), το οποίο έχει χτιστεί πάνω στις ακολουθίες Observations των trajectories εκπαίδευσης. Κάθε observation αντιστοιχίζεται σε ένα μοναδικό αναγνωριστικό (Id) και η τοπολογία του δέντρου κωδικοποιεί τις ακολουθίες αυτών των Ids. Οι ενέργειες που εκτελέστηκαν σε κάθε σημείο δεν αποτελούν μέρος της τοπολογίας του δέντρου, αλλά αποθηκεύονται ως metadata στους κόμβους με την μορφή κατανομών συχνότητας, ένα επιπρόσθετο χαρακτηριστικό που προσδίδει αξιοπιστία στην

δομή. Συνεπώς, το Δέντρο είναι στην θέση να απαντήσει στο ερώτημα: « Πότε βρέθηκε ξανά ο agent στην τρέχουσα κατάσταση και ποιες ενέργειες ακολούθησαν τότε;». Η επιλογή του Δέντρου δεν έγινε τυχαία, καθώς στηρίχτηκε στο χαρακτηριστικό των περιβάλλον όπως αυτό του WebShop, όπου τα trajectories περιέχουν συστηματικά επαναλαμβανόμενα τμήματα. Ως αποτέλεσμα αυτού, ο agent επιστρέφει στις ίδιες σελίδες π.χ. μεταξύ αποτελεσμάτων αναζήτησης και σελίδας προϊόντος και συνεπώς το Δέντρο Επιθημάτων είναι η δομή που επιτρέπει την αποδοτική ανεύρεση κάθε τέτοιου επαναλαμβανομένου τμήματος, ανεξαρτήτως της θέσης του agent εντός ενός trajectory.

Ως δεύτερο ετερογενές συστατικό δρα το μικρό γλωσσικό μοντέλο που χρησιμοποιήθηκε, το Qwen2.5-3B,το οποίο μέσω fine-tuning εκπαιδεύτηκε να λαμβάνει hints από το GST και φέρει την ευθύνη της τελικής απόφασης για το ποια ενέργεια θα εκτελεστεί. Η κατανομή των ρόλων αυτών, συνεπώς, αποτυπώνει την αρχική σχεδιαστική δομή πως η συμβολική δομή πλαισιώνει το μοντέλο λειτουργώντας βοηθητικά και δεν το αντικαθιστά. Αυτό, γιατί το Δέντρο προσφέρει μια ντετερμινιστική, άμεσα αναπαραγώγιμη ανάκληση του ιστορικού, ενώ το μοντέλο μέσω της κατανόησης της τρέχουσας κατάστασης και του στόχου που του έχει δοθεί σε φυσική γλώσσα, αλλά και της δυνατότητας να γενικεύσει καταστάσεις χωρίς ακριβές ιστορικό προηγούμενο, καθιστώντας τον συνδυασμό των δύο μια πολύ ισχυρή προσέγγιση. Μάλιστα, η υπόθεση πως οι ρόλοι αυτοί λειτουργούν συμπληρωματικά επιβεβαιώνεται και εμπειρικά στο Κεφάλαιο 5.

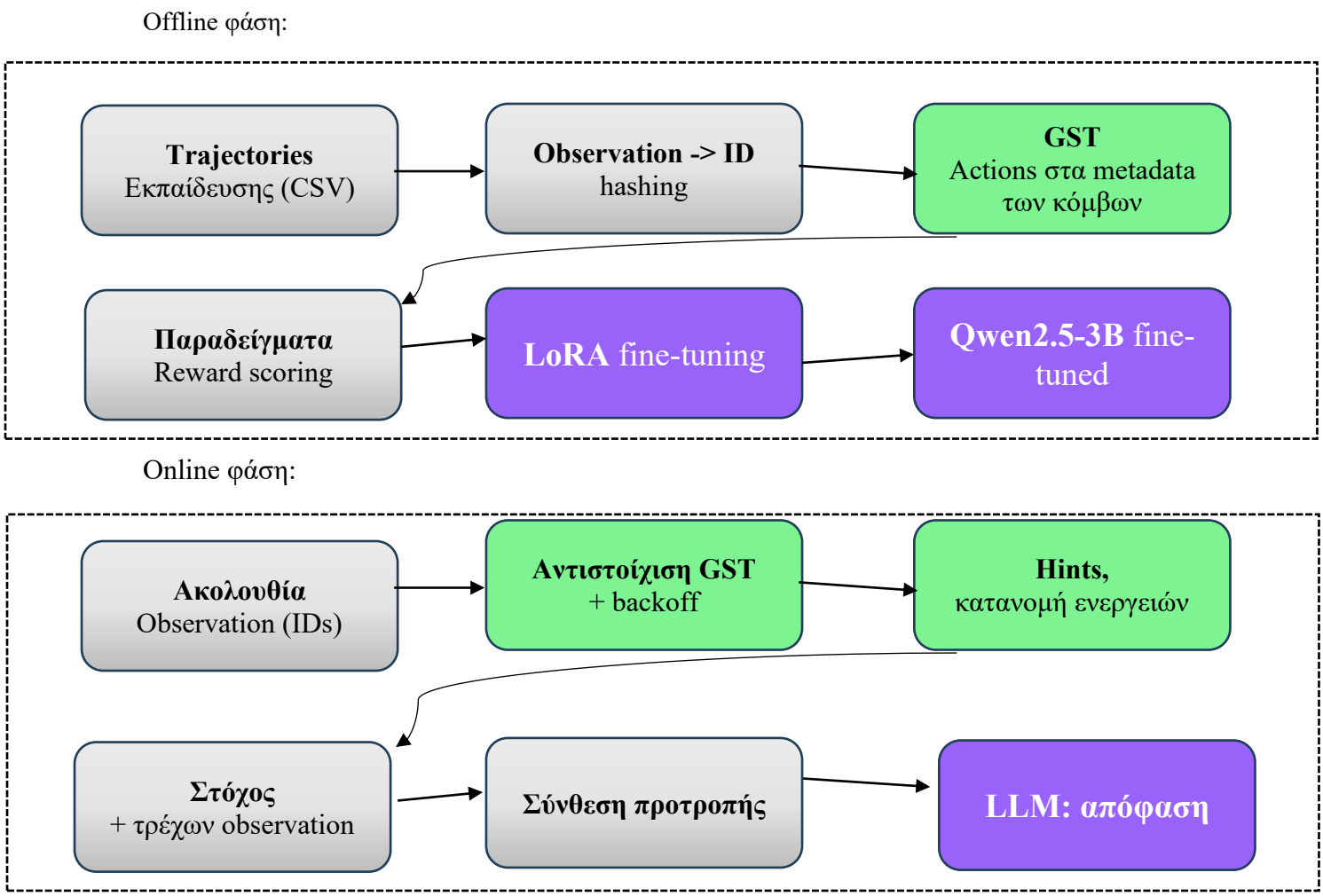
Κατόπιν, η λειτουργία της δομής οργανώνεται σε δύο φάσεις. Στην offline φάση, όπου εκτελείται μια και μόνο φορά: τα trajectories μετασχηματίζονται σε ακολουθίες Observation Ids και εισάγονται στο Δέντρο, με τις αντίστοιχες ενέργειες να καταχωρούνται στα metadata των κόμβων, αλλά και παράγονται από την ίδια πηγή τα παραδείγματα του fine-tuning βάσει μιας βαθμολόγησης που στηρίζεται στην ανταμοιβή. Αντίστοιχα, στην online φάση, σε κάθε βήμα ενός νέου task εκτελείται ένας κύκλος τριών σταδίων:

1) Η ακολουθία των Observations που έχει συναντήσει ο agent έως τώρα αντιστοιχίζεται στο Δέντρο, με μηχανισμό υποχώρησης (backoff) που μικραίνει σταδιακά το εξεταζόμενο επίθημα έως ότου βρεθεί αντιστοίχιση και ανακτάται κατανομή των ενεργειών που ακολούθησαν ιστορικά.
2) Τα hints αυτά σε συνδυασμό με την τρέχουσα παρατήρηση και τον στόχο συντίθεται σε μια προτροπή.
3) Το μοντέλο στέλνει την επόμενη ενέργεια

Ακόμα, κρίσιμη σχεδιαστική απόφαση θεωρείται και η στρατηγική δρομολόγησης. Αντί να υλοποιηθεί μια ιεραρχική προσέγγιση όπου όταν το δέντρο εντοπίζει

ιστορικό προηγούμενο υψηλής βεβαιότητας, η ενέργεια θα εκτελείται απευθείας από αυτό έχοντας το γλωσσικό μοντέλο ως εφεδρεία σε περίπτωση μικρής βεβαιότητας του Δέντρου, επιλέχθηκε μια στρατηγική always-decide. Το μοντέλο, δηλαδή, καλείται σε κάθε βήμα ανεξαιρέτως, με τα hints να λειτουργούν αποκλειστικά συμβουλευτικά. Η επιλογή αυτή έγινε έχοντας ως δεδομένο την πληροφορία που φέρει κάθε συστατικό. Δηλαδή το ιστορικό ναι μεν συνοψίζει το παρελθόν, δεν βλέπει όμως ούτε την τρέχουσα σελίδα ούτε τον στόχο του task. Την πληροφορία αυτή τη διαθέτει μόνο το γλωσσικό μοντέλο, συνεπώς αν μια ενέργεια εκτελεστεί απευθείας από το δέντρο, λαμβάνεται χωρίς καμία γνώση του παρόντος, χρησιμοποιώντας μόνο ιστορικές ενδείξεις (μοτίβα), γι αυτό στο προτεινόμενο σύστημα το μοντέλο αποφασίζει σε κάθε βήμα.

Η αρχιτεκτονική απεικονίζεται στην εικόνα 4.1. Οι ενότητες που ακολουθούν αναλύουν την κατασκευή του Δέντρου (4.2), τον μηχανισμό υποδείξεων (4.3), την ενσωμάτωση του μοντέλου και την δρομολόγηση (4.4) και τον σχεδιασμό των δεδομένων προσαρμογής(4.5). Οι λεπτομέρειες υλοποίησης και τα πειραματικά αποτελέσματα παρατίθενται στο Κεφάλαιο 5.



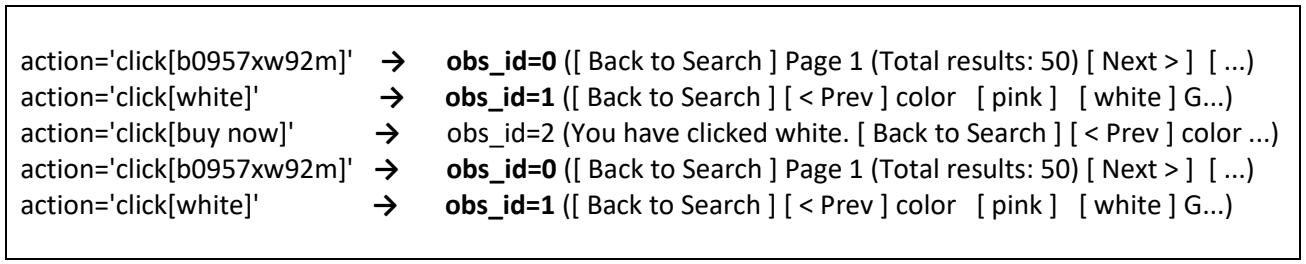

**Εικόνα 4.1:** Συνολική αρχιτεκτονική του υβριδικού συστήματος GST-LLM

## *4.2 Κατασκευή του GST*

Η ορθή κατασκευή του Δέντρου Επιθημάτων αποτελεί ένα από τα πιο κρίσιμα μέρη αυτής της προσέγγισης, καθώς είναι αυτή η δομή που θα καθοδηγήσει το μοντέλο στην ορθή λήψη απόφασης. Η κατασκευή του ξεκινά από τα trajectories του συνόλου εκπαίδευσης, όπου κάθε trajectory είναι μια ακολουθία ζευγών (action, observation). Συνεπώς, το πρώτο βήμα είναι η αντιστοίχιση κάθε παρατήρησης σε ένα ακέραιο αναγνωριστικό (ID): το κείμενο του observation κατακερματίζεται (SHA-256) και την πρώτη φορά που εμφανίζεται λαμβάνει τον επόμενο διαθέσιμο ακέραιο. Με αυτόν τον τρόπο έχουμε επιτύχει το να λαμβάνουν το ίδιο ID, ανεξαρτήτως trajectory, δύο πανομοιότυπες παρατηρήσεις, π.χ. η ίδια σελίδα προϊόντος στην οποία επιστρέφει ο agent κάποια στιγμή. Αξίζει να σημειωθεί πως η αντιστοίχιση αυτή είναι εξωτερική του δέντρου, δηλαδή το Δέντρο χειρίζεται μόνο ακολουθίες ακεραίων.

Στη συνέχεια, πριν από την εισαγωγή, το τελευταίο ζεύγος κάθε επεισοδίου αφαιρείται, γεγονός που αποτελεί σχεδιαστική επιλογή, καθώς στο περιβάλλον WebShop, το τελικό βήμα κάθε επιτυχημένου trajectory είναι κατά κανόνα η ενέργεια: click[buy now], η οποία ως στόχος πρόβλεψης είναι τετριμμένη, δηλαδή θα ήταν πάντοτε η ίδια, που συνεπώς δεν προσφέρει χρήσιμη πληροφορία.

Κατόπιν, κάθε trajectory εισάγεται στο δέντρο ως ακολουθία Observation Ids, ακολουθούμενη από έναν μοναδικό διαχωριστικό όρο ανά trajectory (ένας αρνητικός ακέραιος), ο οποίος εξασφαλίζει πως κανένα επίθημα ενός trajectory δεν συγχέεται με κάποιο άλλης, την τυπική δηλαδή τεχνική με την οποία ένα Δέντρο Επιθημάτων γενικεύεται σε πολλαπλές ακολουθίες. Ακόμα, τα actions δεν συμμετέχουν στην τοπολογία άμεσα: καταχωρίζονται ως metadata στους κόμβους που διασχίζει η ακολουθία, με τη μορφή κατανομών συχνότητας.




**Εικόνα 4.2:** Πραγματική γραμμή των δεδομένων πριν τον διαχωρισμό σε επεισόδια. Μετά το click[buy now] ξεκινά δεύτερο επεισόδιο του ίδιου task, γι' αυτό τα observations 0 και 1 επανεμφανίζονται.

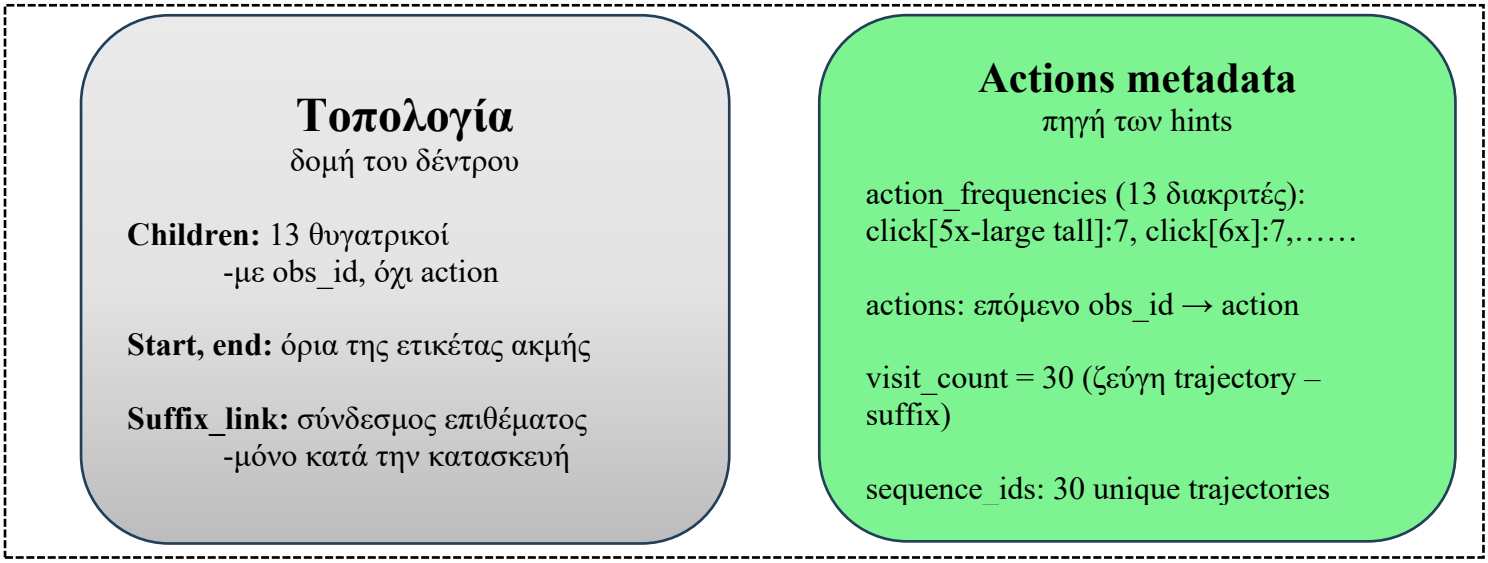
Η παραπάνω εικόνα 4.2 παρουσιάζει μια πραγματική γραμμή του συνόλου εκπαίδευσης, όπως ήταν πριν από τον διαχωρισμό σε επεισόδια. Τα observation IDs 0 και 1 επανεμφανίζονται στις θέσεις 1-2 και 4-5, επειδή μετά το click[buy now] η ίδια γραμμή περιέχει δεύτερο επεισόδιο του ίδιου task. Στο τελικό pipeline κάθε γραμμή χωρίζεται στα επεισόδιά της και από κάθε επεισόδιο αφαιρείται το τελικό βήμα (Ενότητα 5.2), ενώ η ένωση των επεισοδίων ήταν μάλιστα πηγή διαρροής εντός συμφραζομένων. Το παράδειγμα δείχνει πάντως και τη γενική ιδιότητα που αξιοποιεί το GST πως επαναλαμβανόμενα τμήματα ακολουθιών, όπως οι επιστροφές σε ίδιες σελίδες αποτελεσμάτων και προϊόντων, εντοπίζονται και ομαδοποιούνται αυτόματα, ανεξαρτήτως θέσης ή trajectory.

Στη συνέχεια, ακολουθεί η εσωτερική δομή του δέντρου: τι αποθηκεύεται σε κάθε κόμβο και πώς αυτό κατασκευάζεται:

**Περιεχόμενο Κόμβων**

Κάθε κόμβος του Δέντρου φέρει δύο ειδών πληροφορίες: η πρώτη είναι τοπολογική και ταυτίζεται με τον κλασικό μηχανισμό του αλγορίθμου κατασκευής, δηλαδή τα όρια της ετικέτας ακμής πάνω στη συνενωμένη ακολουθία , τους συνδέσμους επιθημάτων (suffix links) που χρησιμοποιούνται μόνο κατά την κατασκευή, και τους θυγατρικούς κόμβους με κλειδί, όπως προαναφέρθηκε, το Observation ID. Η δεύτερη πληροφορία είναι τα metadata των actions, που προστίθενται σε δεύτερο πέρασμα μετά την εισαγωγή κάθε trajectory και αποτελούνται από: μια κατανομή συχνότητας των actions που εκτελέστηκαν στον κόμβο από όλα τα trajectories που το διασχίζουν (action_frequencies), την αντιστοίχιση κάθε action με το Observation ID στην οποία οδήγησε (actions), έναν κατάλογο των διακριτών trajectories που διέρχονται από τον κόμβο (sequence_ids) και τέλος έναν μετρητή επισκέψεων (visit_count). Αξίζει να σημειωθεί επίσης πως το Δέντρο δεν αποθηκεύονται ούτε rewards, ούτε συνέχειες trajectories ως αυτοτελή αντικείμενα, καθώς οι ιστορικές συνέχειες ανασυγκροτούνται κατ' απαίτηση από την τοπολογία και τα metadata.

Κόμβος Δέντρου (πραγματικό παράδειγμα: obs_id=20)




**Εικόνα 4.3:** Περιεχόμενο ενός πραγματικού κόμβου του Δέντρου

**Σημασιολογία του μετρητή επισκέψεων (visit_count)**
Ο μετρητής αυτός απαιτεί προσεκτικό ορισμό, καθώς δεν μετρά διακριτά trajectories, αλλά αυξάνεται μια φορά για κάθε επίθημα κάθε trajectory που διέρχεται από τον κόμβο. Στη ρίζα π.χ. ο μετρητής ισούται με 1228, το άθροισμα δηλαδή του μήκους όλων των trajectories, ενώ τα διακριτά επεισόδια είναι 305. Στο πιο πολυσύχναστο εσωτερικό κόμβο του δέντρου, δηλαδή ο μετρητής (30) υπερεκτιμά τα διακριτά trajectories (30) κατά συντελεστή 1,00. Η ιδιότητα αυτή είναι δομική συνέπεια του σχήματος καταμέτρησης και επανεξετάζεται στην παρακάτω ενότητα, όπου ο μετρητής συμμετέχει στον υπολογισμό της βεβαιότητας των hints.

**Κατασκευή**
Ο αλγόριθμος που επιλέχθηκε με βάση τον οποίο να κατασκευαστεί το Δέντρο είναι ο Ukkonen, στην γενικευμένο μορφή που παρουσιάστηκε σε προαναφερθέν Κεφάλαιο, με πλήρη υλοποίηση του κλασικού μηχανισμού. Δηλαδή ένα ενεργό σημείο (active point), μια βελτιστοποίηση καθόδου skip/count, suffix links και τους τρεις κανόνες επέκτασης. Η επιλογή αυτή εγγυάται την ορθή κλιμάκωση πως ο αλγόριθμος επιτυγχάνει αποσβεσμένο γραμμικό χρόνο ως προς το συνολικό μήκος των ακολουθιών, έναντι του τετραγωνικού της απλοϊκής εισαγωγής. Στην πράξη, για το σύνολο εκπαίδευσης των 305 επεισοδίων από 209 tasks (923 observation IDs συνολικά), η κατασκευή ολοκληρώνεται σε δευτερόλεπτα και παράγει δέντρο 1544 κόμβων, εκ των οποίων 1228 φύλλα. Συνεπώς

το υπολογιστικό κόστος του συστήματος δεν εντοπίζεται στη δομή του Δέντρου αλλά στο νευρωνικό σκέλος του γλωσσικού μοντέλου.

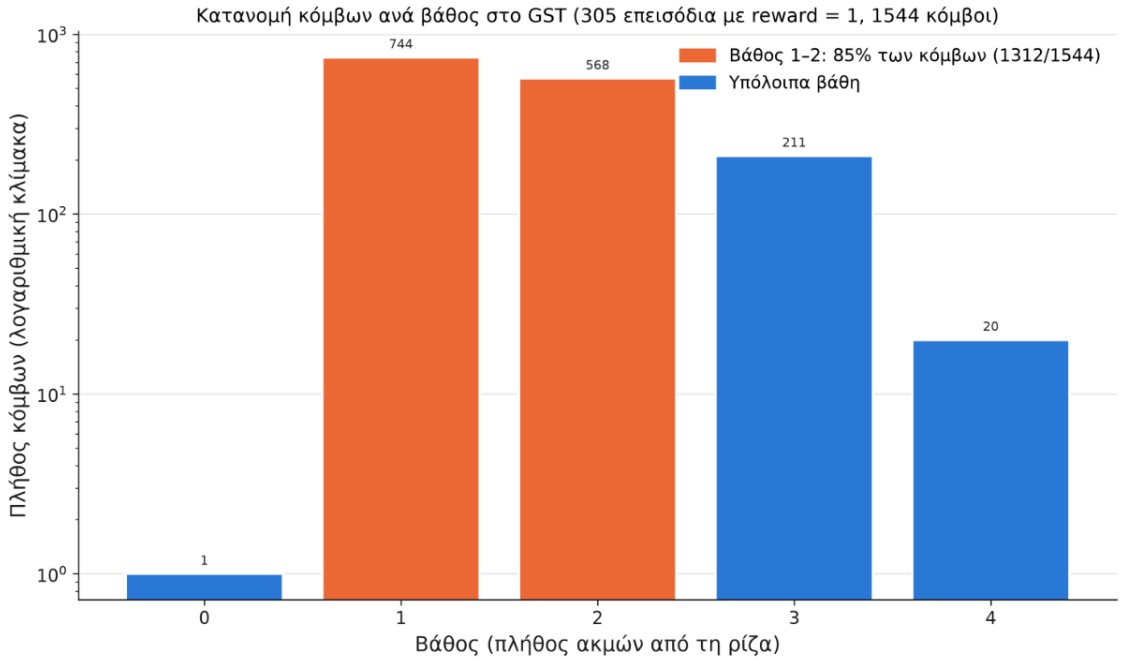
**Ανατομία του κατασκευασμένου Δέντρου**

Ο πίνακας 4.1 συνοψίζει τα στατιστικά του Δέντρου που χρησιμοποιείται σε όλα τα πειράματα:

| **Μέγεθος** | **Τιμή** |
|---|---|
| Επεισόδια (tasks) | 305 (209) |
| Observations | 923 |
| Unique Observations | 439 |
| Κόμβοι | 1544 |
| Φύλλα | 1228 |
| Actions | 923 |
| Unique Actions | 292 |
| Μέσο visit_count | ≈2,20 |

**Πίνακας 4.1:** Στατιστικά του Δέντρου

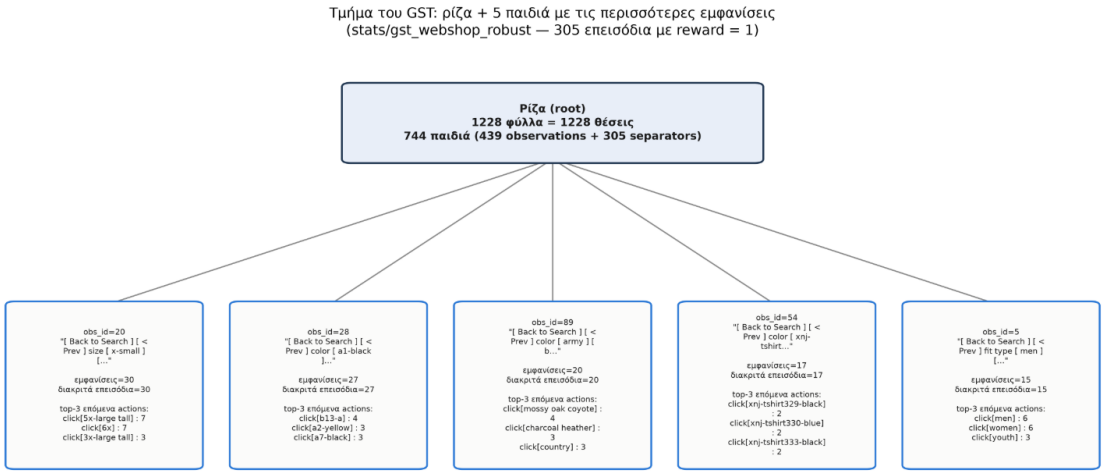
Η εικόνα 4.4 παρακάτω απεικονίζει την κατανομή των κόμβων ανά βάθος, η οποία αποκαλύπτει το χαρακτηριστικότερο γνώρισμα του δέντρου, πως το 85% των κόμβων (1312 από 1544) βρίσκεται σε βάθος 1 έως 2, ενώ πέραν του βάθους 10 απομένουν μονοψήφιοι κόμβοι ανά επίπεδο. Το δέντρο είναι, με λίγα λόγια, ρηχό και ευρύ, δηλαδή πλούσιο σε μικρά κοινά μοτίβα, φτωχό σε μακρά. Το εύρημα αυτό συνάδει με το γνώρισμα των δεδομένων πως είναι αραιά, καθώς σχεδόν τα μισά observations (49,7%) εμφανίζονται μία μόνο φορά σε ολόκληρο το σύνολο εκπαίδευσης, ενώ ακόμα έχει και δύο άμεσες συνέπειες αρχιτεκτονικές που θα αναλυθούν παρακάτω περαιτέρω: πρώτον, η αντιστοίχιση μακρών επιθημάτων θα αποτυγχάνει συχνά, καθιστώντας τον μηχανισμό υποχώρησης όχι μια απλή εξαίρεση αλλά κανόνα λειτουργίας και δεύτερον, τα hints που ανακτώνται θα αφορούν συχνά ελάχιστες ιστορικές μαρτυρίες, γεγονός που καθιστά αναγκαία την ύπαρξη ενός «κριτή», δηλαδή το γλωσσικό μοντέλο.




**Εικόνα 4.4:** Κατανομή κόμβων ανά βάθος

**Ενδεικτικό απόσπασμα**

Η εικόνα 4.5 στη συνέχεια παρουσιάζει τους πέντε κόμβους του πρώτου επιπέδου με τον υψηλότερο μετρητή επισκέψεων. Σε αυτό το σημείο αξίζει να παρατηρηθούν δύο σημαντικές λεπτομέρειες. Πρώτον, όλες οι ακμές πρώτου επιπέδου που εμφανίζονται έχουν μήκος ετικέτας 1, η διακλάδωση, δηλαδή, συμβαίνει αμέσως, χωρίς μακρά αδιακλάδωτα τμήματα κοντά στη ρίζα. Δεύτερον, παρατηρούμε πως και οι πέντε κόμβοι αντιστοιχούν σε σελίδες επιλογής χαρακτηριστικών (χρώμα, μέγεθος κτλ.), δηλαδή οι πιο συχνές περιοχές του δέντρου είναι ακριβώς εκείνες που πολλαπλές, παρόμοιες ενέργειες ανταγωνίζονται (π.χ. στον κορυφαίο κόμβο καταγράφονται 13 διακριτές ενέργειες). Η παρατήρηση αυτή προοικονομεί ένα κεντρικό εύρημα της ανάλυσης σφαλμάτων του Κεφαλαίου 5, πως τα λάθη επιλογής χαρακτηριστικού αναδεικνύονται στην δυσκολότερη κατηγορία του προβλήματος.




**Εικόνα 4.5:** Οι πέντε κόμβοι του πρώτου επιπέδου με το υψηλότερο visit_count.

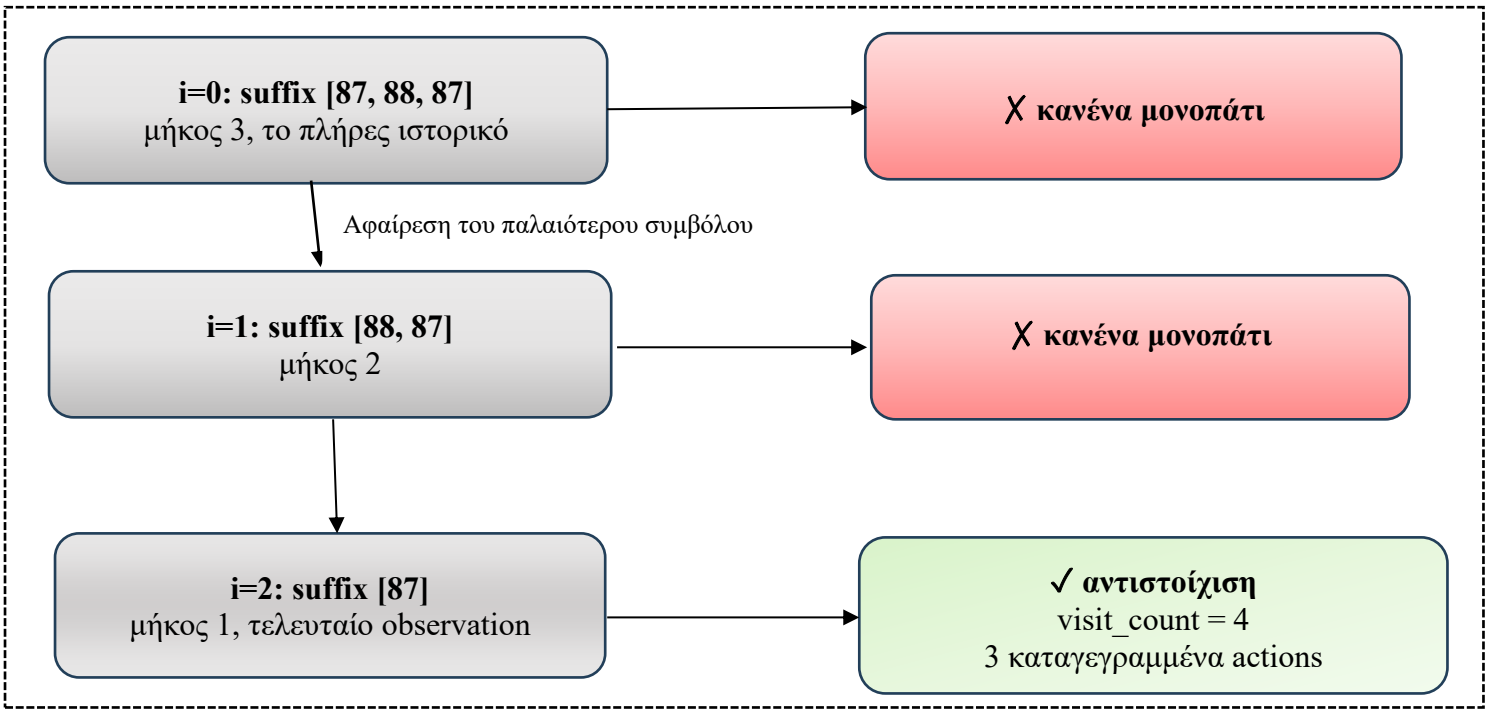
## *4.3 Μηχανισμός ιστορικών υποδείξεων (hints)*

**Αντιστοίχιση**
Η ανάκτηση των hints ξεκινά από την ακολουθία των observation ids που έχει συγκεντρώσει ο agent στο τρέχον επίπεδο και στη συνέχεια πραγματοποιείται η αντιστοίχιση στο δέντρο που είναι ακριβής: η αναζήτηση διασχίζει το δέντρο συγκρίνοντας observation ids ένα προς ένα και μία μόνο διαφορά στο επίθημα συνεπάγεται πλήρη αποτυχία της αντιστοίχισης, δηλαδή κάνουμε exact match. Η αυστηρότητα αυτή είναι σχεδιαστική απόφαση , καθώς κληρονομεί την ντετερμινιστική επαληθεύσιμη φύση της δομής, με αντίτιμο την ευαισθησία σε παραλλαγές, δηλαδή δύο σχεδόν πανομοιότυπες σελίδες λαμβάνουν διαφορετικά ids και συνεπώς δεν ταυτίζονται ποτέ.

**Υποχώρηση (backoff)**
Δεδομένης της αραιότητας που εξηγήθηκε παραπάνω , δηλαδή έχουμε ένα ρηχό δέντρο με σπάνια μακρά κοινά μονοπάτια, η αποτυχία αντιστοίχισης του πλήρους ιστορικού δεν είναι εξαίρεση αλλά ο κανόνας και ο μηχανισμός υποχώρησης την χειρίζεται ως εξής: η

αναζήτηση εκκινά από το μακρύτερο επίθημα (ολόκληρο το ιστορικό) και στη συνέχεια σε κάθε αποτυχία αφαιρεί το παλαιότερο σύμβολο, δοκιμάζοντας διαδοχικά μικρότερα επιθήματα έως το μήκος 1. Η διαδικασία τερματίζει στην πρώτη έγκυρη αντιστοίχιση, δηλαδή στο μακρύτερο ιστορικό προηγούμενο που εμφανίζεται στο δέντρο. Από το σημείο αντιστοίχισης, οι ιστορικές ενέργειες και η συχνότητά τους ανακτώνται από το σύνολο των εμφανίσεων (occurrences) που αντιστοιχούν στο επίθημα και για κάθε εμφάνιση λαμβάνεται η ενέργεια που ακολούθησε στη συγκεκριμένη θέση, ώστε η υπόδειξη να αντανακλά πιστά την πραγματική συνέχεια κάθε ιστορικής εμφάνισης. Αν κανένα suffix δεν αντιστοιχηθεί, το σύστημα επιστρέφει κενή υπόδειξη και το μοντέλο αποφασίζει χωρίς ιστορική καθοδήγηση.

Η Εικόνα 4.6 απεικονίζει τον μηχανισμό σε ενδεικτικό παράδειγμα: από το ιστορικό [87, 88, 87], τα επιθήματα μήκους 3 και 2 δεν υπάρχουν στο δέντρο, και μόνο το μονοσύμβολο επίθημα [87] αντιστοιχίζεται, από το οποίο ανακτώνται οι ιστορικά καταγεγραμμένες ενέργειες.




**Εικόνα 4.6:** Ενδεικτικό παράδειγμα του μηχανισμού υποχώρησης (backoff): διαδοχική βράχυνση του επιθήματος έως την πρώτη έγκυρη αντιστοίχιση (εδώ στο μήκος 1).

**Βαθμολόγηση υποδείξεων**

Στον κόμβο της αντιστοίχισης, κάθε καταγεγραμμένο action λαμβάνει βαθμό ίσο με το γινόμενο δύο όρων: της σχετικής συχνότητάς της στον κόμβο (πόσο συχνά, δηλαδή, επιλέχθηκε εκεί ιστορικά) και ενός συντελεστή βεβαιότητας που αυξάνεται με το visit_count του κόμβου και φτάνει έως την τιμή 1 όταν ο μετρητής φτάσει το 10, ώστε οι κατανομές που κατέχουν ελάχιστες «μαρτυρίες» να υποβαθμίζονται.. Σε αυτό το σημείο επιστρέφουμε στο προαναφερθέν ζήτημα του γεγονότος πως επειδή το visit_count καταμετρά ζεύγη (trajectory, suffix) και όχι διακριτά trajectories, ο κορεσμός επέρχεται στην πράξη με λιγότερες ανεξάρτητες μαρτυρίες απ' όσες υποδηλώνει η ονομαστική τιμή 10, π.χ. στον πολυσύχναστο κόμβο που αναλύσαμε προηγουμένως με συντελεστή υπερεκτίμησης 2.24. Το γεγονός καταγράφεται ως γνωστός περιορισμός βαθμονόμησης της φόρμουλας, δηλαδή οι βαθμοί υποδείξεων είναι συγκρίσιμοι μεταξύ τους εντός κόμβου, αλλά δεν πρέπει να ερμηνεύονται ως καλά βαθμονομημένες πιθανότητες.

**Το φίλτρο ελάχιστων επισκέψεων**

Ο μηχανισμός αυτός προβλέπει προαιρετικό κατώφλι που αποκλείει αντιστοιχίσεις σε κόμβους με λίγες επισκέψεις. Στην αναπτυγμένη διαμόρφωση, το κατώφλι είναι απενεργοποιημένο (δηλαδή έχει τιμή 1, την οποία ικανοποιεί κάθε κόμβος εξ ορισμού). Η ενεργοποίησή του σε μια δοκιμή με τιμή 2 επιδείνωσε την ακρίβεια, καθώς σε ένα τόσο αραιό Δέντρο η απόρριψη των κόμβων που «μαρτυρούν» ελάχιστη πληροφορία ισοδυναμεί συχνά με απόρριψη κάθε διαθέσιμου hint. Η επιλογή αυτή συνάδει με την σκέψη πως μια αδύναμη υπόδειξη έχει αξία μόνο εφόσον ο τελικός κριτής, που είναι το μοντέλο, μπορεί να την σταθμίσει.

**Οι δύο παράλληλες αναζητήσεις**

Για λόγους πληρότητας και επειδή η ιεραρχική δρομολόγηση αποτελεί βασική γραμμή σύγκρισης στο Κεφάλαιο 5, σημειώνεται πως το σύστημα υλοποιεί δύο παραλλαγές της αναζήτησης με backoff, με διαφορετική σημασιολογία τερματισμού όπως φαίνεται και στον πίνακα 4.2 παρακάτω. Η παραλλαγή των hints που περιεγράφηκε ανωτέρω, τερματίζει στην πρώτη έγκυρη αντιστοίχιση, καθώς έχει ως προτεραιότητα το μακρύτερο suffix, διότι η ποιοτική στάθμιση θα γίνει ούτως η άλλως από το μοντέλο. Η παραλλαγή της ιεραρχικής δρομολόγησης, αντιθέτως, εξετάζει όλα τα μήκη suffixes και επιλέγει καθολικά τη μέγιστη βεβαιότητα, την οποία και υποβάλλει σε κατώφλι αποδοχής, δηλαδή όταν το Δέντρο πρόκειται να αποφασίσει αυτόνομα,χωρίς τον έλεγχο του μοντέλου, η αυστηρότερη επιλογή είναι δικαιολογημένη. Αξίζει σε αυτό το σημείο να παρατηρήσουμε πως η δεύτερη παραλλαγή, λόγω του συντελεστή βεβαιότητας τείνει να ευνοεί βραχέα επιθήματα με υψηλά visit_counts, δηλαδή δημοφιλή αλλά φτωχές αντιστοιχίσεις.

| | Υποδείξεις (always-decide & ιεραρχική) | Ιεραρχική απόφαση |
|---|---|---|
| **Σειρά εξέτασης** | Μακρύτερο → βραχύτερο επίθημα | Όλα τα μήκη |
| **Τερματισμός** | Πρώτη έγκυρη αντιστοίχιση | Καθολικό μέγιστο βεβαιότητας |
| **Προτεραιότητα** | Μήκος suffix | Ύψος βεβαιότητάς |
| **Πύλη αποδοχής** | -- (κενή υπόδειξη στο πλήρες miss) | Κατώφλι βεβαιότητας |
| **Ρόλος εξόδου** | Συμβουλευτικός (κρίνει το LLM) | Δεσμευτικός (αποφασίζει το Δέντρο |

**Πίνακας 4.2:** Οι δύο παραλλαγές αναζήτησης με υποχώρηση

## *4.4 Ενσωμάτωση LLM και Always-Decide approach*

**Σύνθεση της προτροπής**
Η επικοινωνία του Δέντρου με το γλωσσικό μοντέλο γίνεται μέσω μιας ενιαίας προτροπής που συγκεντρώνει όλη τη διαθέσιμη πληροφορία, δηλαδή το goal του task, το τρέχον observation, τις πλέον πρόσφατες ενέργειες που πραγματοποίησε ο agent και τα ιστορικά hints. Τα hints παρατίθενται ως ένας κατάλογος ζευγών action-βεβαιότητας ταξινομημένος κατά φθίνουσα βεβαιότητα και περιορισμένος στις πέντε κορυφαίες από όλες. Το όριο των πέντε αυτών βεβαιοτήτων είναι αποκλειστικά συμβατικό στοιχείο μορφοποίησης της προτροπής και όχι ιδιότητα του μηχανισμού υποδείξεων, καθώς το αντικείμενο της υπόδειξης μπορεί να περιέχει πολύ περισσότερες όπως π.χ. ο κόμβος που εξετάστηκε προηγουμένως που φέρει 13 στο σύνολο, αλλά οι πέντε μόνο κορυφαίες μεταφέρονται στο κείμενο. Ομοίως, από το ιστορικό των actions περιλαμβάνονται μόνο οι πέντε τελευταίες. Στο τέλος, η προτροπή κλείνει με ρητή οδηγία επιλογής μίας μόνο ενέργειας, ενώ η μορφή της απεικονίζεται πλήρως, σε πραγματικό παράδειγμα, στην Εικόνα 4.8.

**Το προεπιλεγμένο μήνυμα συστήματος**
Στο μοντέλο διαβιβάζεται ένα και μόνο μήνυμα, το οποίο περιέχει ολόκληρη την προτροπή , συνεπώς η υλοποίηση δεν ορίζει δικό της μήνυμα συστήματος. Κατά τη μορφοποίηση της εισόδου σύμφωνα με το πρότυπο συνομιλίας του μοντέλου, το Qwen2.5-3B εισάγει αυτόματα το δικό του προεπιλεγμένο μήνυμα συστήματος («You

are Qwen, created by Alibaba Cloud. You are a helpful assistant.»). Το κείμενο αυτό δεν αποτελεί σχεδιαστική επιλογή της παρούσας εργασίας, αλλά προεπιλογή του ίδιου του μοντέλου και συνεπώς το αναφέρουμε εδώ για λόγους ακρίβειας και αναπαραγωγιμότητας, ώστε η περιγραφή της εισόδου να αντιστοιχεί ακριβώς σε αυτό που λαμβάνει το μοντέλο.

**Εξαγωγή της ενέργειας**

Η απόκριση του μοντέλου αναλύεται συντακτικά ώστε να εξαχθεί έγκυρη ενέργεια, δηλαδή αναζητείται πρώτα ρητή δήλωση ενέργειας και αν αυτή απουσιάζει, τότε το πρώτο έγκυρο συντακτικά μοτίβο ενέργειας οπουδήποτε στην απόκριση. Η προσέγγιση αυτή προσδίδει ανθεκτικότητα έναντι αποκρίσεων με περιττά λόγια, χωρίς να απαιτεί από το μοντέλο να διατηρήσει ένα αυστηρό μορφότυπο.

**Δρομολόγηση: Ιεραρχική έναντι always-decide**

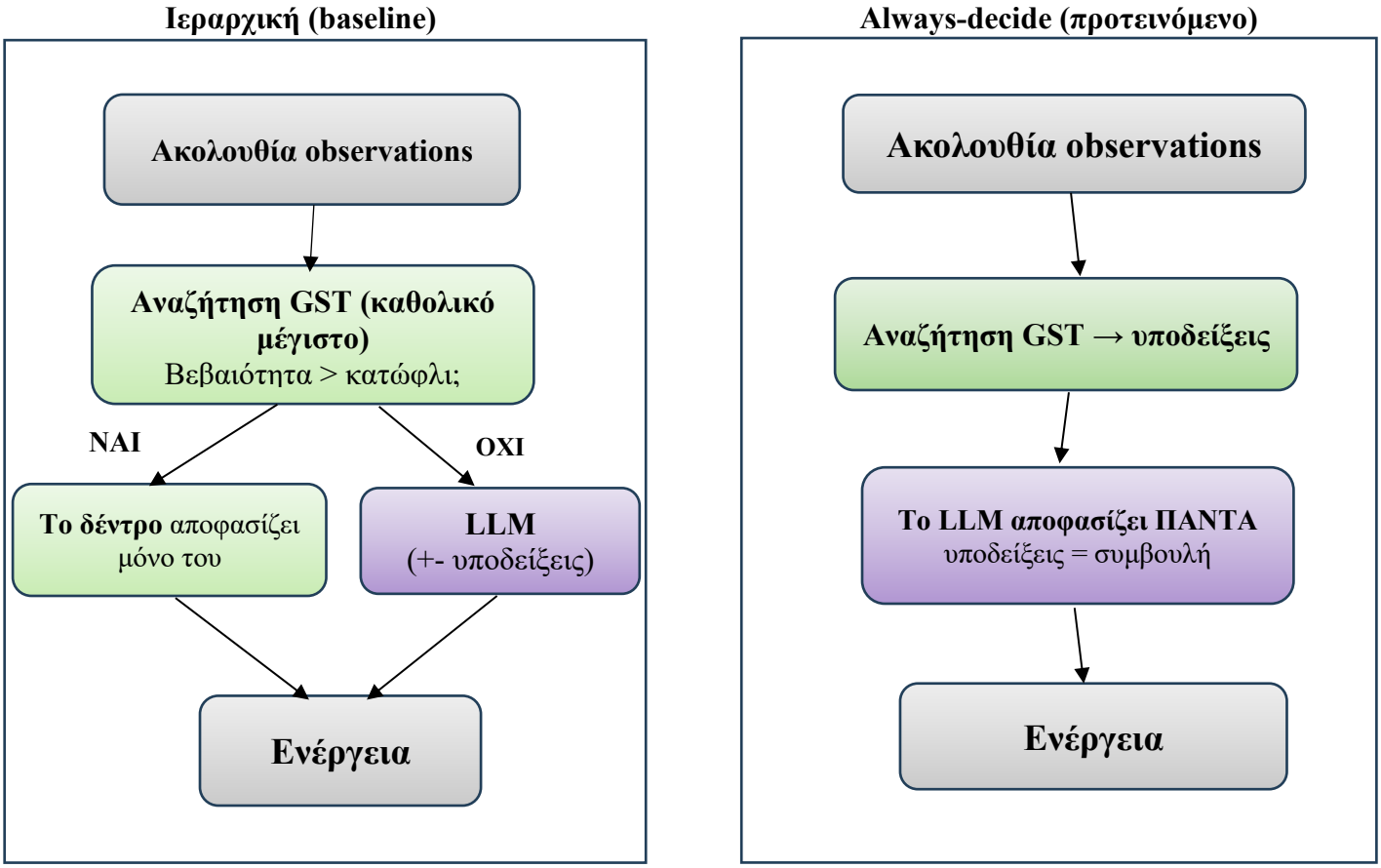
Η κεντρική σχεδιαστική απόφαση του συστήματος αφορά το ποια δομή λαμβάνει την τελική απόφαση και το σύστημα υλοποιεί δύο εναλλακτικές στρατηγικές (Εικόνα 4.7). Στην ιεραρχική στρατηγική, που χρησιμεύει ως baseline σύγκριση, η αναζήτηση στο Δέντρο εκτελείται πρώτη. Σε περίπτωση που εντοπίσει αντιστοίχιση της οποίας η βεβαιότητα υπερβαίνει ένα κατώφλι, η αντίστοιχη ενέργεια εκτελείται απευθείας και το μοντέλο δεν καλείται καθόλου. Μόνο όταν το Δέντρο αποτύχει να αποφασίσει μόνο του με επαρκή βεβαιότητα, η απόφαση μεταβιβάζεται στο μοντέλο. Στη στρατηγική always-decide αντιθέτως, την οποία και υιοθετεί το προτεινόμενο σύστημα, το Δέντρο δεν αποφασίζει ποτέ αυτόνομα, αλλά η αναζήτηση χρησιμοποιείται αποκλειστικά για την παραγωγή των υποδείξεων και στη συνέχεια το μοντέλο καλείται σε κάθε βήμα ανεξαιρέτως είτε όταν υπάρχουν είτε όταν δεν υπάρχουν υποδείξεις.

Η διαφορά αυτή μεταξύ των δύο προσεγγίσεων είναι θεμελιώδης, καθώς στην ιεραρχική στρατηγική υπάρχει μια ολόκληρη κατηγορία αποφάσεων που λαμβάνονται χωρίς καμία αναφορά στην τρέχουσα παρατήρηση ή goal, αφού η αυτόνομη απόφαση του δέντρου στηρίζεται μόνο στην στατιστική του παρελθόντος. Αντιθέτως, η στρατηγική του always-decide εξαλείφει αυτή την κατηγορία, καθώς κάθε απόφαση περνά υποχρεωτικά από το μόνο συστατικό που αντιλαμβάνεται το παρόν, δηλαδή το LLM. Εδώ εντοπίζεται και η ουσιαστική διαφοροποίηση από προσεγγίσεις που παρεμβαίνουν στο προπαρασκευαστικό στάδιο, όπως το LCoW, ενώ η παρέμβαση του παρόντος συστήματος δεν αφορά το τι βλέπει το μοντέλο, αλλά διασφαλίζει πως το μοντέλο αποφασίζει πάντα, με το Δέντρο να έχει έναν ρόλο συμβουλευτικό και όχι να αποτελεί υποκατάστατό του.

**Η φύση των υποδείξεων ως συμβουλής**

Η επιλογή του always-decide για το υπάρχον σύστημα ανέδειξε ένα εύλογο ερώτημα: αν το μοντέλο αγνοεί τις υποδείξεις, τότε αυτές τι νόημα έχουν; Η απάντηση σε αυτό το ερώτημα αναδεικνύεται από το ίδιο το παράδειγμα που συνοδεύει αυτό το Κεφάλαιο και που όπως φαίνεται στην εικόνα 4.8, το μη fine-tuned μοντέλο καλείται να αποφασίσει έχοντας ως κορυφαία υπόδειξη την ορθή ενέργεια (click[3x-large], με βεβαιότητα 0,20). Ωστόσο, επιλέγει διαφορετική εσφαλμένη ενέργεια. Το παράδειγμα αυτό που σκόπιμα παρατίθεται ως αποτυχία και όχι ως επιτυχία, καθιστά χειροπιαστό ένα κεντρικό εύρημα:

πως το μοντέλο δεν αξιοποιεί αξιόπιστα τις υποδείξεις χωρίς προσαρμογή. Η παρατήρηση αυτή δεν αναιρεί την στρατηγική απόφαση, αλλά την αιτιολογεί και συνιστά άμεση γέφυρα προς την ανάγκη για fine-tuning, καθώς οι υποδείξεις αποκτούν αξία μόνο όταν το μοντέλο εκπαιδευτεί να τις σταθμίζει σε συνδυασμό με την τρέχουσα κατάσταση.




**Εικόνα 4.7:** Σχηματικό διάγραμμα των δύο στρατηγικών δρομολόγησης

Goal: Find me machine wash, wash cold women's fashion hoodies &
sweatshirts for dry clean, tumble dry with color: navy blue, and
size: 3x-large, and price lower than 60.00 dollars

Current Observation:
[ Back to Search ] [ < Prev ]
size: [ small ] [ medium ] ... [ 3x-large ] [ 4x-large ]
color: [ black ] ... [ navy blue ] ... [ sand camo ] [ vintage camo ]
Expression Tees Pug Life Funny Thug Life Unisex Adult Hoodie
Price: $36.99 to $48.99 [...]

Previous Actions:
click[3x-large]
click[navy blue]

Suggested Actions (from historical patterns):
- click[3x-large] (confidence: 0.20)
- click[sand camo] (confidence: 0.10)
- click[medium] (confidence: 0.10)

Based on the goal, current state, and suggested actions,
what is the best next action? Respond with the action only!

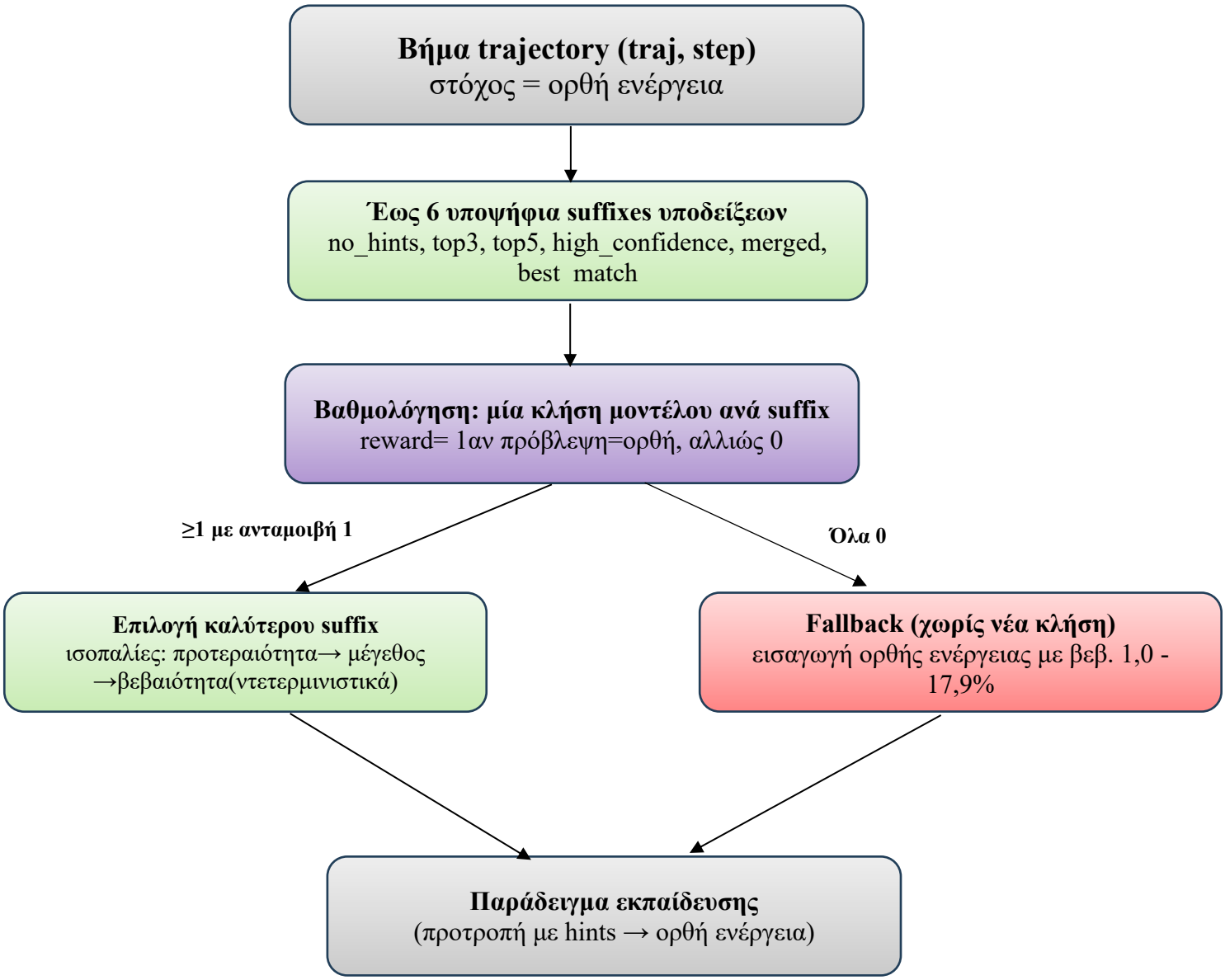
---

**Απόκριση μοντέλου: click[navy blue]** ✗

**Ορθή ενέργεια: click[3x-large]**

**Εικόνα 4.8:** Πραγματική προτροπή και εσφαλμένη απόκριση του μη fine-tuned μοντέλου, παρά την ορθή κορυφαία υπόδειξη.

## *4.5 Σχεδιασμός Δεδομένων προσαρμογής*




**Εικόνα 4.9:** Παραγωγή παραδείγματος εκπαίδευσης με βαθμολόγηση ανταμοιβής και μηχανισμό fallback για τα βήματα μηδενικής ανταμοιβής

Όπως τεκμηριώθηκε στην ενότητα 4.4, το μη fine-tuned μοντέλο δεν αξιοποιεί αξιόπιστα τα ιστορικά hints και γι' αυτό τον σκοπό κρίνεται απαραίτητο το fine-tuning ώστε να διδάξει στο μοντέλο πώς να σταθμίζει τα hints σε συνδυασμό με την τρέχουσα κατάσταση. Η ενότητα αυτή περιγράφει τον σχεδιασμό των δεδομένων προσαρμογής, ένα ζήτημα που στο παρόν σύστημα δεν είναι απλή προετοιμασία αλλά ενσωματώνει μια ουσιαστική σχεδιαστική ιδέα.

**Δομή ενός παραδείγματος**

Κάθε παράδειγμα εκπαίδευσης αντιστοιχεί σε ένα βήμα ενός trajectory του συνόλου εκπαίδευσης και έχει τη μορφή ζεύγους προτροπής-απόκρισης. Δηλαδή, η είσοδος είναι μια προτροπή της ίδιας μορφής με εκείνη της ενότητας 4.4 (στόχος, τρέχον observation, πρόσφατες ενέργειες, υποδείξεις), ενώ ο στόχος είναι πάντοτε η πραγματική ορθή ενέργεια του βήματος. Αξίζει να σημειωθεί πως ο στόχος δεν είναι ποτέ κείμενο παραγόμενο από το μοντέλο, αντιθέτως είναι πάντα η τεκμηριωμένη ενέργεια του trajectory. Με τον τρόπο αυτό, το μοντέλο εκπαιδεύεται να παράγει τη σωστή ενέργεια δεδομένου ενός συνόλου υποδείξεων και όχι να αναπαράγει τις υποδείξεις αυτές.

**Η κεντρική ιδέα: επιλογή συνόλου υποδείξεων με βαθμολόγηση ανταμοιβής**
Ένα ερώτημα που αγείρεται είναι το ποιες υποδείξεις να συνοδεύουν κάθε παράδειγμα. Μια αφελής επιλογή θα ήταν να χρησιμοποιείται πάντα το ίδιο είδος υποδείξεων, ωστόσο το παρόν σύστημα υιοθετεί μια πιο εκλεπτυσμένη προσέγγιση. Για κάθε βήμα παράγονται έως έξι εναλλακτικά σύνολα υποδείξεων, από την πλήρη απουσία υποδείξεων, έως διάφορες παραλλαγές (οι τρεις ή πέντε κορυφαίες, μόνο οι υψηλής βεβαιότητας, η ένωση όλων των αντιστοιχίσεων). Για την παραγωγή αυτών, χρησιμοποιείται μια πλουσιότερη εκδοχή του μηχανισμού αναζήτησης, η οποία σε αντίθεση με τον μηχανισμό της ενότητας 4.3 που τερματίζει στην πρώτη αντιστοίχιση, συλλέγει όλους τους αντιστοιχισμένους κόμβους σε όλα τα μήκη suffixes.

Κατόπιν, κάθε εναλλακτικό σύνολο αξιολογείται εμπειρικά: το μοντέλο καλείται μία φορά ανά σύνολο και λαμβάνει δυαδική ανταμοιβή, δηλαδή λαμβάνει την μονάδα εάν προβλέψει την ορθή ενέργεια, αλλιώς μηδέν. Στη συνέχεια, ως τελικό σύνολο του παραδείγματος επιλέγεται εκείνο που πέτυχε ανταμοιβή, με ντετερμινιστική επίλυση ισοπαλιών (κατά προτεραιότητα παραλλαγής, έπειτα μικρότερο μέγεθος, έπειτα υψηλότερη μέση βεβαιότητα). Η λογική πίσω από αυτό είναι ουσιώδης, καθώς με αυτόν τον τρόπο το σύστημα δεν επιλέγει το «πλουσιότερο» σύνολο, αλλά εκείνο που αποδεδειγμένα οδηγεί το μοντέλο στη σωστή απόφαση, μια αρχή που παραπέμπει στη φιλοσοφία επιλογής βάσει ανταμοιβής του LCoW, εφαρμοσμένη όμως στην επιλογή υποδείξεων αντί στην αναδιατύπωση observations. Η συνολική διαδικασία απεικονίζεται στην Εικόνα 4.9 όπως φαίνεται παραπάνω.

Σημειώνεται, επίσης, μια ιδιότητα της σχεδίασης με συνέπειες για την αξιολόγηση. Επειδή το Δέντρο κατασκευάζεται από τα ίδια επεισόδια που χρησιμοποιούνται για την εκπαίδευση, κάθε βήμα εκπαίδευσης «βλέπει» στο Δέντρο τη δική του μετάβαση. Στο παραγόμενο σύνολο, η σωστή ενέργεια περιλαμβάνεται πάντοτε στα hints του νικητή (703/703 παραδείγματα με hints) ή δεν υπάρχουν hints. Το μοντέλο δεν εκτίθεται ποτέ σε περίπτωση όπου υπάρχουν hints που δεν περιέχουν τη σωστή ενέργεια. Η ιδιότητα αυτή αποδείχθηκε καθοριστική και εξετάζεται πειραματικά στην Ενότητα 5.9.

**Αντιμετώπιση των βημάτων μηδενικής ανταμοιβής**

Ένα αρκετά ενδιαφέρον σημείο που παρατηρήθηκε είναι πως σε ένα σημαντικό ποσοστό των βημάτων (17,9% των αρχικών παραδειγμάτων) κανένα από τα εναλλακτικά σύνολα δεν οδήγησε το μοντέλο στη σωστή ενέργεια. Συνεπώς, για να αντιμετωπίσουμε τις περιπτώσεις αυτές, το σύστημα δεν επαναλαμβάνει την κλήση του μοντέλου, αντ' αυτού κατασκευάζει ντετερμινιστικά ένα συνθετικό σύνολο, εισάγοντας την ορθή ενέργεια με βεβαιότητα 1,0 ως πρώτη υπόδειξη, μαζί με τις υποδείξεις του πλουσιότερου διαθέσιμου συνόλου όπου υπήρχε, και το χρησιμοποιεί ως παράδειγμα εκπαίδευσης. Στην πράξη, επειδή τα πανομοιότυπα σύνολα συγχωνεύονται, σε 144 από τα 165 τέτοια παραδείγματα η μοναδική υπόδειξη ήταν η ίδια η ορθή ενέργεια. Η επιλογή αυτή διασφαλίζει πως ακόμη και τα πιο δύσκολα βήματα συνεισφέρουν σήμα εκπαίδευσης, το μοντέλο δηλαδή βλέπει τη σωστή ενέργεια συσχετισμένη με ένα εύλογο σύνολο, αλλά εγείρει παράλληλα και ένα ζήτημα ισορροπίας, το οποίο αντιμετωπίζεται στο επόμενο βήμα.

**Σύνθεση του συνόλου**

Στο τελικό σύνολο προσαρμογής, το ποσοστό των συνθετικών (fallback) παραδειγμάτων είναι 17,9%, δηλαδή σαφής μειοψηφία και καθώς τα συνθετικά παραδείγματα δεν κυριαρχούν στο σύνολο, δεν κρίθηκε αναγκαία κάποια ισοστάθμιση, δηλαδή το σύνολο των 923 παραδειγμάτων χρησιμοποιήθηκε ως έχει για την εκπαίδευση. Η πλειονότητα των παραδειγμάτων προέρχεται από πραγματικές υποδείξεις του Δέντρου (κυρίως τα σύνολα top-3 και merged), ενώ τα παραδείγματα χωρίς υποδείξεις (no_hints) αποτελούν περίπου το ένα τέταρτο.

| Παραλλαγή συνόλου | Πλήθος(923) |
|---|---|
| Μηδενικής ανταμοιβής (συνθετικό) | 165 (17,9%) |
| top3 | 516 (55,9%) |
| merged_all_suffixes | 22 (2,4%) |
| Χωρίς υποδείξεις | 220 (23,8%) |
| **Σύνολο** | **923** |

**Πίνακας 4.3:** Σύνθεση του συνόλου προσαρμογής (923 παραδείγματα)

## *4.6 Σύνοψη Σχεδιαστικών Αποφάσεων*

Η αρχιτεκτονική που παρουσιάστηκε στις προηγούμενες ενότητες προέκυψε από μια σειρά διακριτών σχεδιαστικών αποφάσεων, καθεμία από τις οποίες συνιστά επιλογή έναντι μιας εύλογης εναλλακτικής. Συνεπώς, στην ενότητα αυτή θα συγκεντρώσουμε όλες αυτές τις αποφάσεις, ώστε η λογική του συστήματος να αποτυπωθεί συνολικά από ένα σημείο (Πίνακας 4.4). Κοινός άξονας όλων είναι η θεμελιώδης αρχή που διατυπώθηκε στην ενότητα 4.1 πως το Δέντρο πλαισιώνει το γλωσσικό μοντέλο χωρίς να το υποκαθιστά.

Ακόμα, αξίζει να σημειωθεί πως δύο από τις αποφάσεις αυτές είναι πλέον καθοριστικές. Η πρώτη είναι η κωδικοποίηση των observations στην τοπολογία του δέντρου, καθώς ορίζει το θεμελιώδες ερώτημα που απαντά το Δέντρο: «πότε ξαναβρέθηκε ο agent σε αυτήν την κατάσταση;» και καθορίζει, επίσης, ολόκληρο τον μηχανισμό ανάκτησης. Ενώ η δεύτερη είναι η στρατηγική του always-decide, καθώς το να έχουμε το μοντέλο να αποφασίζει σε κάθε βήμα με το Δέντρο να έχει αυστηρά συμβουλευτικό ρόλο είναι το γεγονός που τοποθετεί την παρέμβαση του συστήματος στο στάδιο της απόφασης και το διαφοροποιεί από προσεγγίσεις που παρεμβαίνουν στο προπαρασκευαστικό στάδιο, δηλαδή στην επεξεργασία των observations. Οι υπόλοιπες αποφάσεις, δηλαδή ο μηχανισμός υποχώρησης, η επιλογή συνόλου υποδείξεων με βαθμολόγηση ανταμοιβής, αλλά και το φιλτράρισμα σε επιτυχημένα επεισόδια αποτελούν μια εξειδίκευση αυτών των δύο κεντρικών επιλογών.

Τέλος, είναι σημαντικό να σημειωθεί πως αρκετές από τις αποφάσεις αυτές επικυρώνονται εμπειρικά στο Κεφάλαιο 5, όπου αναλύεται η υπεροχή του always-decide έναντι της ιεραρχικής δρομολόγησης, η αναγκαιότητα του fine-tuning και ο συμπληρωματικός ρόλος των δύο συστατικών (Δέντρου και γλωσσικού μοντέλου), οι οποίες αποφάσεις δεν παρουσιάζονται ως αξιώματα αλλά ως ισχυρισμοί που ελέγχονται πειραματικά.

| Απόφαση | Εναλλακτική | Αιτιολόγηση |
|---|---|---|
| Κωδικοποίηση παρατηρήσεων στην τοπολογία, ενέργειες ως metadata | Κωδικοποίηση ακολουθιών ενεργειών | Το ερώτημα-κλειδί είναι «πού βρέθηκα ξανά;» και οι ενέργειες είναι η απάντηση που ανασύρεται, όχι το κλειδί αναζήτησης |
| Ακριβής αντιστοίχιση observation IDs | Ασαφής/σημασιολογική αντιστοίχιση | Διατηρεί τον ντετερμινιστικό και επαληθεύσιμο χαρακτήρα της Δομής |
| Υποχώρηση με τερματισμό στην πρώτη αντιστοίχιση για τις υποδείξεις | Καθολικό μέγιστο βεβαιότητας | Προτεραιότητα στο μακρύτερο σύνολο, η ποιοτική στάθμιση γίνεται από το μοντέλο |
| Απενεργοποίηση φίλτρου ελάχιστων επισκέψεων | Ενεργό φίλτρο | Σε αραιό Δέντρο, η απόρριψη κόμβων με λίγες πληροφορίες ισοδυναμεί με απόρριψη κάθε υπόδειξης. Σε διερευνητική δοκιμή επιδείνωσε την ακρίβεια (Ενότητα 5.7) |
| Στρατηγική always-decide | Ιεραρχική Δρομολόγηση | Το Δέντρο δεν αντιλαμβάνεται το παρόν, κάθε απόφαση πρέπει να περνά από το μόνο συστατικό που βλέπει την τρέχουσα κατάσταση |
| Επιλογή συνόλου υποδείξεων με βαθμολόγηση ανταμοιβής | Σταθερό είδος υποδείξεων | Επιλέγεται εμπειρικά το σύνολο που οδηγεί το μοντέλο στη σωστή απόφαση, όχι το θεωρητικά «πλουσιότερο» |
| Δυαδική ανταμοιβή | Βαθμωτή ανταμοιβή | Στο WebShop δεν υπάρχει «σχεδόν σωστή» ενέργεια, η επιλογή είναι διακριτή |
| Χρήση όλων των παραδειγμάτων χωρίς ισοστάθμιση | Υποδειγματοληψία των συνθετικών | Το ποσοστό των συνθετικών (fallback) παραδειγμάτων (17,9%) είναι μειοψηφικό, οπότε δεν υπάρχει ανισορροπία προς διόρθωση |
| Χρήση μόνο πλήρως επιτυχημένων επεισοδίων (reward=1) | Συμπερίληψη και μερικώς επιτυχημένων (reward<1) | Ένα μερικώς αποτυχημένο trajectory περιέχει ενέργειες που δεν οδηγούν στην ορθή λύση, άρα δεν αποτελεί έγκυρο ground truth — η χρήση της θα τιμωρούσε το μοντέλο ακόμη κι όταν επιλέγει την ορθή ενέργεια |

**Πίνακας 4.4:** Η λογική του Συστήματος

# 5

# *Υλοποίηση και Αποτελέσματα*

## *5.1 Υπολογιστικό περιβάλλον και αναπαραγωγιμότητα*

Στο κεφάλαιο αυτό θα μεταβούμε από τον σχεδιασμό στην υλοποίηση, συνεπώς στη συνέχεια θα περιγραφεί το υπολογιστικό περιβάλλον, οι ρυθμίσεις εκπαίδευσης, η πειραματική διάταξη και το σύνολο των αποτελεσμάτων που ελέγχουν τους σχεδιαστικούς ισχυρισμούς του Κεφαλαίου 4. Κύρια προτεραιότητα όλης της παρουσίασης αποτελεί η αναπαραγωγιμότητα, δηλαδή κάθε αριθμητικό αποτέλεσμα που αναφέρεται προέρχεται από εκτελέσεις με καθορισμένο seed τυχαιότητας και τεκμηριωμένη διαμόρφωση.

**Υπολογιστική υποδομή**
Το σύνολο των πειραμάτων εκτελέστηκε στον διακομιστή DGX του Πανεπιστημίου Πατρών, σε μία κάρτα γραφικών NVIDIA A100-SXM4-40GB, όπου σύμφωνα με την πολιτική χρήσης πόρων του συστήματος, κάθε πείραμα δεσμεύει το πολύ μία κάρτα. Το Λειτουργικό Σύστημα είναι Ubuntu 22.04.5 LTS και η υλοποίηση αναπτύχθηκε σε περιβάλλον Python με βιβλιοθήκες: PyTorch (έκδοση 2.5.1, με ενσωματωμένο περιβάλλον εκτέλεσης CUDA 12.1), Transformers (5.14.1) και PEFT (0.19.1). Το βασικό μοντέλο είναι το Qwen2.5-3B-Instruct, το οποίο προσαρμόστηκε μέσω LoRa χωρίς

τροποποίηση των βαρών του πυρήνα, μια προσέγγιση που καθιστά την εκπαίδευση εφικτή σε μία μόνο κάρτα, καθώς εκπαιδεύεται μόνο ένα μικρό κλάσμα των παραμέτρων.

**Αναπαραγωγιμότητα και έλεγχος τυχαιότητας**

Θεμελιώδης μεθοδολογική αρχή της παρούσας εργασίας είναι πως ότι αναφέρεται ως αποτέλεσμα θα πρέπει να είναι αναπαραγώγιμο. Για τον σκοπό αυτό, όλα τα στάδια του pipeline που εμπεριέχουν τυχαιότητα αρχικοποιούνται με σταθερό seed (=42, envs), δηλαδή ο διαχωρισμός των δεδομένων, η κατασκευή του συνόλου προσαρμογής, η εκπαίδευση και η αξιολόγηση. Επίσης, σε κάθε στάδιο, το seed εφαρμόζεται στις σχετικές γεννήτριες τυχαίων αριθμών (των βιβλιοθηκών random, NumPy, PyTorch, αλλά και της γεννήτριας της κάρτας γραφικών), ενώ η αποκωδικοποίηση του μοντέλου εκτελείται με χαμηλό temperature δειγματοληψίας (0,3) ώστε να περιορίζεται η μεταβλητότητα.

Σε αυτό το σημείο κρίνεται απαραίτητη μια ακριβής διατύπωση, πως ο έλεγχος της τυχαιότητας γίνεται σε επίπεδο κατάστασης των γεννητριών, δηλαδή δεν ενεργοποιούνται οι αυστηροί ντετερμινιστικοί αλγόριθμοι της PyTorch (use_deterministic_algorithms). Παρ' όλα αυτά, ο ντετερμινισμός των αποτελεσμάτων επαληθεύτηκε εμπειρικά έχοντας δύο διαδοχικές εκτελέσεις της ίδιας διαμόρφωσης οι οποίες παρήγαγαν πανομοιότυπα, ανά βήμα, αποτελέσματα (με ταυτόσημα αρχεία εξόδου). Η αναπαραγωγιμότητα, δηλαδή, δεν στηρίζεται σε θεωρητική εγγύηση της βιβλιοθήκης, αλλά σε εμπειρικά επιβεβαίωση επί της συγκεκριμένης υποδομής μίας κάρτας A100. Η διασφάλιση αυτή είναι κρίσιμη, δεδομένου πως η κεντρική συνεισφορά της εργασίας είναι μια εσωτερική, ελεγχόμενη σύγκριση διαμορφώσεων, εννοώντας πως η εγκυρότητα της σύγκρισης εξαρτάται από το ότι οι παρατηρούμενες διαφορές οφείλονται στις υπό εξέταση μεταβλητές και όχι σε τυχαία διακύμανση.

**Δομή του Κεφαλαίου**

Στη συνέχεια, στην ενότητα 5.2 θα περιγραφεί ο διαχωρισμός των δεδομένων σε επίπεδο task και ο τρόπος με τον οποίο αποτρέπεται η διαρροή της πληροφορίας. Κατόπιν στην ενότητα 5.3 θα παρατεθούν οι ρυθμίσεις της προσαρμογής LoRA, ενώ στην 5.4 ορίζεται η πειραματική διάταξη. Τέλος, η ενότητα 5.5 παρουσιάζει τα κύρια αποτελέσματα με τα διαστήματα εμπιστοσύνης τους, ενώ η 5.6 την ανάλυση σφαλμάτων και η 5.7 τα συμπληρωματικά ευρήματα.

## *5.2 Διαχωρισμός Δεδομένων και Αποτροπή Διαρροής*

**Προέλευση και μορφή των δεδομένων**
Τα δεδομένα της παρούσας εργασίας προέρχονται από τις συλλογές trajectories του WebShop, που διατίθενται το πλαίσιο της μεθοδολογίας expert iteration του LCoW (Lee et al., 2025). Αυτά αποτελούνται από δύο διαδοχικές επαναλήψεις (iterations) συλλογής επί κοινού συνόλου tasks, όπου κάθε επανάληψη παράγει βελτιωμένα trajectories για τα ίδια υποκείμενα προβλήματα. Κάθε γραμμή των αρχείων αντιστοιχεί σε ένα βήμα trajectory και περιλαμβάνει το id του trajectory, την εκτελεσθείσα ενέργεια και το πλήρες κείμενο observation στην ακατέργαστη μορφή text_rich του περιβάλλοντος. Σημειώνεται πως τα observations αυτά δεν έχουν υποστεί την πλαισίωση (contextualization) του LCoW, πρόκειται δηλαδή για ακατέργαστο κείμενο του περιβάλλοντος, γεγονός συμβατό με τη θέση της παρούσας εργασίας πως η παρέμβαση δε γίνεται στην αναπαράσταση των παρατηρήσεων αλλά στο στάδιο της απόφασης.

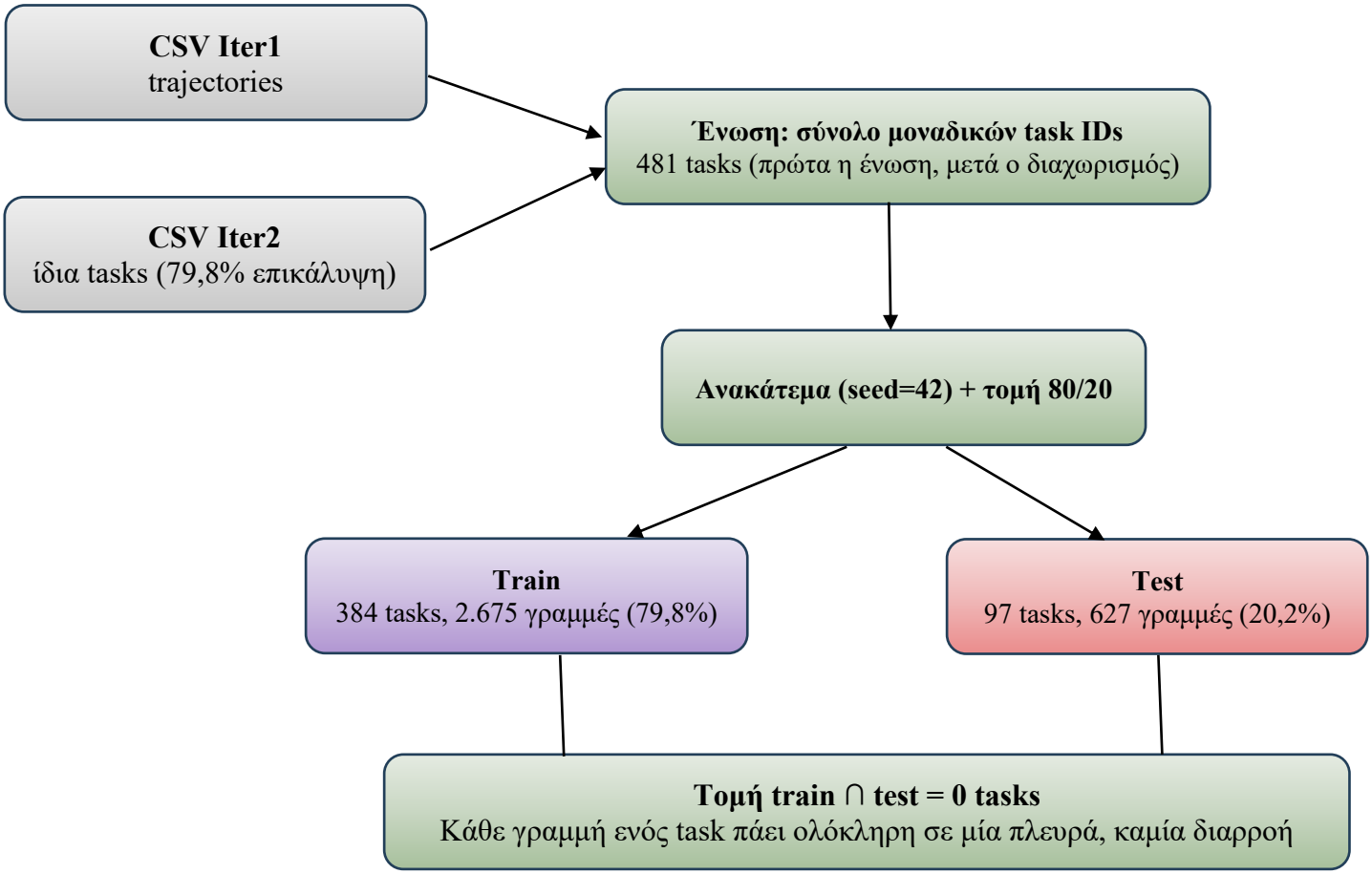
**Το πρόβλημα της επικάλυψης**
Ο ορθός διαχωρισμός των δεδομένων σε ένα σύνολο εκπαίδευσης και συνόλου ελέγχου αποτελεί ίσως μια από τις πιο κρίσιμες διαδικασίες, καθώς επηρεάζει άμεσα την αξιοπιστία του συστήματος. Στο πλαίσιο του WebShop, το ζήτημα αυτό αποκτά ιδιαίτερη διάσταση λόγω ενός χαρακτηριστικού των διαθέσιμων δεδομένων: οι δύο πηγές trajectories προέρχονται από διαδοχικές επαναλήψεις (iterations) της ίδιας διαδικασίας συλλογής των ίδιων tasks,έχοντας ως αποτέλεσμα υψηλή επικάλυψη σε επίπεδο task μεταξύ τους. Η επικάλυψη αυτή δεν αποτελεί σφάλμα των δεδομένων, αλλά είναι εγγενές χαρακτηριστικό της μεθόδου συλλογής τους (επαναληπτική βελτίωση επί κοινού συνόλου tasks). Ωστόσο, αν αυτή η επικάλυψη αγνοηθεί κατά τον διαχωρισμό των δεδομένων, μπορεί να οδηγήσει σε διαρροή πληροφορίας, καθώς αν τα trajectories του ίδιου task κατέληγαν και στα δύο σύνολα, το μοντέλο θα αξιολογούνταν σε tasks που έχει ήδη δει κατά την εκπαίδευση παρέχοντας ψευδώς αισιόδοξα αποτελέσματα.

**Γιατί οι απλοί διαχωρισμοί αποτυγχάνουν**
Είναι σημαντικό να εξεταστεί σε αυτό το σημείο το γιατί οι τρεις προφανείς εναλλακτικές είναι ακατάλληλες εδώ. Ένας διαχωρισμός σε επίπεδο γραμμής, θα κατένειμε βήματα του ίδιου trajectory σε διαφορετικές πλευρές, προφανώς άτοπο. Αντίστοιχα, ένας διαχωρισμός σε επίπεδο trajectory που αποτελεί τη σύνηθη πρακτική, θα ήταν επαρκής για δεδομένα μίας επανάληψης, ωστόσο εδώ θα έπρεπε δύο trajectories του ίδιου task (ένα από κάθε επανάληψη) να βρεθούν σε αντίθετες πλευρές, μεταφέροντας ακέραιη την επικάλυψη στον διαχωρισμό. Τέλος, ο ανεξάρτητος διαχωρισμός κάθε πηγής θα επέτρεπε σε ένα task να είναι εκπαιδευτικό στη μία πηγή και ελεγκτικό στην άλλη. Συνεπώς, όπως παρατηρείτε, κοινός παρονομαστής και των τριών αποτυχιών είναι πως αγνοούν πως η μονάδα που πρέπει να παραμείνει αδιαίρετη δεν είναι η γραμμή ούτε το trajectory, αλλά το task.

**Διαχωρισμός σε επίπεδο task**

Η λύση που υιοθετείται είναι ο διαχωρισμός σε επίπεδο task (task-level split) και όχι σε επίπεδο trajectory ή γραμμής. Η διαδικασία αυτή απεικονίζεται στην Εικόνα 5.1 και εκτελείται ως εξής:

Πρώτα, οι δύο πηγές συνενώνονται και συλλέγεται το σύνολο των μοναδικών task ids επί της ένωσης και όχι χωριστά ανά πηγή. Το βήμα αυτό κρίνεται κρίσιμο, καθώς με το να αντιμετωπίζεται η ένωση των tasks ως ενιαία δεξαμενή που διαχωρίζεται μία και μόνο φορά, αποκλείεται εξ ορισμού το ενδεχόμενο ένα task να εμφανιστεί και στις δύο πλευρές. Στη συνέχεια, το σύνολο των tasks αναδιατάσσεται τυχαία (με το σταθερό seed) και κατόπιν διαχωρίζεται σε αναλογία 80/20 σε δύο ξένα μεταξύ τους υποσύνολα. Τέλος, κάθε γραμμή δεδομένων δρομολογείται στην πλευρά όπου ανήκει το task της, ώστε όλα τα trajectories ενός task να καταλήγουν ακέραια στην ίδια πλευρά. Το κρίσιμο βήμα είναι το πρώτο, δηλαδή με την συλλογή των ids επί της ένωσης πριν από οποιονδήποτε διαχωρισμό, το ενδεχόμενο διαρροής αποκλείεται δομικά και όχι μέσω εκ των υστέρων φιλτραρίσματος.




**Εικόνα 5.1:** Διαχωρισμός σε επίπεδο task. Η ένωση των ids τέμνεται σε ξένα σύνολα ώστε κανένα task να μην εμφανίζεται και στις δύο πλευρές

**Επαλήθευση**
Αξίζει να σημειωθεί πως κώδικας υπολογίζει και αναφέρει την τομή των δύο αυτών συνόλων tasks σε κάθε εκτέλεση με ρητή προειδοποίηση σε περίπτωση μη κενής τομής. Ο διαχωρισμός παρήγαγε 481 μοναδικά tasks συνολικά, εκ των οποίων 384 (79,8%) στην εκπαίδευση και 97 (20,2%) στον έλεγχο, με μηδενική τομή, μια ιδιότητα που επαληθεύτηκε τόσο κατά την εκτέλεση όσο και εκ νέου επί των τελικών αρχείων. Σε επίπεδο γραμμών, επίσης, το σύνολο εκπαίδευσης περιλαμβάνει 2.675 βήματα και το σύνολο ελέγχου 627.

**Διαχρωρισμός σε επεισόδια και φιλτραρίσματα ανταμοιβής**
Ένα επιπλέον, κρίσιμο χαρακτηριστικό των δεδομένων είναι ότι πολλαπλά επεισόδια του ίδιου task (ένα από κάθε iteration) αποθηκεύονται ενωμένα σε ένα ενιαίο trajectory. Αν αυτό αγνοηθεί, προκαλεί διαρροή εντός συμφραζομένων (in-context leakage), καθώς η λύση ενός προηγούμενου επεισοδίου εμφανίζεται στο ιστορικό ενεργειών ενός επόμενου για τον ίδιο ακριβώς στόχο. Για την αντιμετώπισή του, κάθε trajectory διασπάται στα επιμέρους επεισόδιά του με βάση το σημείο επανεκκίνησης της αναζήτησης, ώστε κάθε επεισόδιο να αποτελεί αυτοτελή μονάδα και να μην διασχίζει καμία ακολουθία ιστορικού τα όριά του.

Επιπλέον, διατηρούνται μόνο τα επεισόδια με ανταμοιβή ίση με τη μονάδα (πλήρως επιτυχημένα), τόσο για την κατασκευή του Δέντρου και την εκπαίδευση όσο και για την αξιολόγηση, με συνεπές κριτήριο και στις δύο πλευρές. Η επιλογή αυτή είναι μεθοδολογικά αναγκαία, καθώς ένα μερικώς αποτυχημένο trajectory περιέχει ενέργειες που δεν οδηγούν στην ορθή λύση, και συνεπώς δεν αποτελεί έγκυρη ετικέτα αξιολόγησης. Η χρήση της θα «τιμωρούσε» το μοντέλο ακόμη και όταν αυτό επιλέγει την ορθή ως προς τον στόχο ενέργεια.

**Σύνολο αξιολόγησης**
Μετά τον διαχωρισμό σε επεισόδια και το φιλτράρισμα ανταμοιβής, το σύνολο εκπαίδευσης περιλαμβάνει 305 επεισόδια και το τελικό σύνολο αξιολόγησης 45 σημεία ελέγχου από 36 επεισόδια. Η αντιστοίχιση observations σε ids κατασκευάζεται αποκλειστικά από τα δεδομένα εκπαίδευσης, αφού κατά την προετοιμασία του συνόλου ελέγχου, κάθε observation αναζητείται μέσω του hashing της στο λεξικό αυτό και observations με κείμενο που δεν εμφανίστηκε κατά την εκπαίδευση παραλείπονται, αφού δεν διαθέτουν id. Το μέγεθος του τελικού συνόλου είναι, επομένως, φυσική συνέπεια του μηχανισμού αντιστοίχισης exact-matching και όχι επιλογή υποδειγματοληψίας, δηλαδή παρατηρήσεις άγνωστες στο Δέντρο (τυπικά νέες καταχωρίσεις προϊόντων) αποκλείονται. Ο σχετικά μικρός αριθμός σημείων ελέγχου αποτελεί βασικό περιορισμό της μελέτης, ο οποίος συζητείται στην Ενότητα 6.2. Σημειώνεται ότι η εγκυρότητα της εσωτερικής σύγκρισης δεν θίγεται, καθώς και οι πέντε διαμορφώσεις αξιολογούνται στο ταυτόσημο σύνολο των 45 σημείων και ο Πίνακας 5.1 συνοψίζει όλα τα μεγέθη.

| Μέγεθος | Τιμή |
|---|---|
| Μοναδικά tasks (ένωση δύο επαναλήψεων) | 481 |
| Tasks εκπαίδευσης / βήματα | 384 (79,8%) / 2.675 |
| Tasks ελέγχου / βήματα | 97 (20,2%) / 627 |
| Τομή train ∩ test | 0 |
| Επεισόδια εκπαίδευσης (reward=1) | 305 |
| Επεισόδια αξιολόγησης (reward=1) | 36(κύρια ανάλυση)/ 69(πλήρες σύνολο) |
| Σημεία ελέγχου: in-distribution | 45 |
| Σημεία ελέγχου: OOD | 68 |
| Σημεία ελέγχου: πλήρες σύνολο | 113 |

**Πίνακας 5.1:** Σύνοψη των συνόλων δεδομένων

Κατά τον έλεγχο της διαδικασίας εξαγωγής των σημείων ελέγχου εντοπίστηκε ότι το φίλτρο αφαιρούσε σε κάθε επεισόδιο το τελευταίο διατηρημένο βήμα, με την παραδοχή ότι πρόκειται για την ενέργεια buy now. Επειδή η σελίδα του buy now είναι σχεδόν πάντα άγνωστη στο Δέντρο και έχει ήδη αποκλειστεί, το βήμα που αφαιρούνταν ήταν στην πράξη πραγματική ενέργεια. Με διορθωμένο φίλτρο (αφαίρεση μόνο όταν η ενέργεια είναι κυριολεκτικά buy now), το σύνολο ελέγχου επεκτείνεται σε 113 σημεία από 69 επεισόδια. Η επέκταση προσθέτει 68 σημεία, καθένα από διαφορετικό επεισόδιο. Σε όλα τα 68 η σωστή ενέργεια απουσιάζει από τις υποδείξεις, και αυτά αποτελούν το καθεστώς εκτός κατανομής (OOD) της Ενότητας 5.8. Το GST και το fine-tuned μοντέλο παραμένουν αμετάβλητα· επαληθεύτηκε επίσης ότι κανένα νέο σημείο δεν εισάγει διαρροή στην προτροπή (0/68).

## *5.3 Ρυθμίσεις LoRA*

Η προσαρμογή του μοντέλου Qwen2.5-3B Instruct όπως προαναφέρθηκε πραγματοποιήθηκε με την μέθοδο LoRA, της οποίας το θεωρητικό υπόβαθρο παρουσιάστηκε στην ενότητα 2.3. Η παρούσα ενότητα τεκμηριώνει την πλήρη διαμόρφωση της εκπαίδευσης, ώστε το πείραμα να είναι αναπαραγώγιμο στο σύνολό του.

**Διαμόρφωση LoRA**

Οι προσαρμογείς εφαρμόστηκαν με τάξη r = 8 και συντελεστή κλιμάκωσης α = 16, σε όλες τις γραμμικές προβολές κάθε στρώματος του transformer, δηλαδή τις τέσσερις προβολές attention (query, key, value, output) και τις τρεις του δικτύου εμπρόσθιας τροφοδότησης (gate, up, down). Ακόμα, δεν προσαρμόστηκαν τα embeddings ούτε η κεφαλή εξόδου. Η επιλογή της ευρείας κάλυψης, αντί της συνήθους πρακτικής περιορισμού στις προβολές attention, δίνει στην προσαρμογή πρόσβαση και στο υπολογιστικό μονοπάτι του δικτύου εμπρόσθιας τροφοδότησης, διατηρώντας με αυτόν τον τρόπο το εκπαιδεύσιμο μέρος αμελητέο: 14,97 εκατομμύρια παράμετροι (0,48% από τους 3,1 δισεκατομμυρίων του μοντέλου).

**Διαμόρφωση εκπαίδευσης**

Η εκπαίδευση εκτελέστηκε για 3 epochs επί του συνόλου της ενότητας 4.5, το οποίο διαχωρίστηκε στρωματοποιημένα σε 831 παραδείγματα εκπαίδευσης και 92 επικύρωσης. Ακόμα, χρησιμοποιήθηκε ρυθμός μάθησης $2x10^{-4}$ με χρονοπρόγραμμα συνημιτόνου και προθέρμανση 10%, δραστικό μέγεθος δέσμης 16 ( 4 ανά συσκευή x 4 βήματα συσσώρευσης κλίσεων), μέγιστο μήκος ακολουθίας 1.024 tokens και αριθμητική ακρίβεια bfloat16, χωρίς κβαντοποίηση του βασικού μοντέλου. Ο βελτιστοποιητής είναι ο προεπιλεγμένος της βιβλιοθήκης Transformers (AdamW), ενώ στο τέλος κάθε epoch αξιολογείται η απώλεια επικύρωσης και διατηρείται το καλύτερο σημείο ελέγχου.

Ακόμα, δύο τεχνικές επιλογές αξίζουν ιδιαίτερης προσοχής με την πρώτη να αφορά την απώλεια να υπολογίζεται μόνο επί των tokens της απόκρισης (της ενέργειας) και όχι επί της προτροπής, δηλαδή το μοντέλο εκπαιδεύεται να παράγει τη σωστή ενέργεια δεδομένης της προτροπής και όχι να αναπαράγει την προτροπή, επιλογή που συγκεντρώνει όλο το εκπαιδευτικό σήμα στο ζητούμενο. Κατόπιν, η δεύτερη επιλογή αφορά τη μη συγχώνευση του προσαρμογέα με τα βάρη του βασικού μοντέλου κατά την αξιολόγηση, καθώς εφαρμόζεται κατά τον υπολογισμό της εμπρόσθιας διέλευσης, διατηρώντας το βασικό μοντέλο ανέπαφο, γεγονός που επιτρέπει την αξιολόγηση του βασικού και fine-tuned μοντέλου από την ίδια εγκατάσταση με μόνη διαφορά τη φόρτωση του προσαρμογέα.

Τέλος, η εκπαίδευση ολοκληρώθηκε σε περίπου 5 λεπτά (299,5 δευτερόλεπτα, 156 βήματα βελτιστοποίησης), με τελική απώλεια εκπαίδευσης 0,0269 και απώλεια επικύρωσης 0,0011. Η πολύ χαμηλή απώλεια επικύρωσης εξηγείται εν μέρει από το ότι ο διαχωρισμός εκπαίδευσης/επικύρωσης γίνεται σε επίπεδο παραδείγματος, με διαστρωμάτωση ανά είδος συνόλου υποδείξεων, με αποτέλεσμα 80 tasks να εμφανίζονται και στα δύο υποσύνολα. Αυτό επηρεάζει μόνο την απώλεια επικύρωσης και την επιλογή του checkpoint, όχι τα αποτελέσματα ελέγχου, καθώς η ουσιαστική δοκιμασία γενίκευσης είναι η αξιολόγηση σε αθέατα tasks που ακολουθεί. Στον Πίνακα 5.2 παρακάτω φαίνεται η πλήρης διαμόρφωση της προσαρμογής LoRA.

| **Παράμετρος** | **Τιμή** |
|---|---|
| Βασικό μοντέλο | Qwen/Qwen2.5-3B-Instruct |
| Τάξη r / συντελεστής α / dropout | 8 / 16/ 0,1 |
| Στοχευόμενες μονάδες | q, k, v, o, gate, up, down (όλες οι γραμμικές προβολές) |
| Εκπαιδεύσιμες παράμετροι | 14,97M (0,48% των 3,1B) |
| Epochs | 3 |
| Ρυθμός μάθησης / χρονοπρόγραμμα | $2x10^{-4}$ / συνημίτονο, προθέρμανση 10% |
| Δραστικό μέγεθος δέσμης | 16 (4x4 συσσώρευση) |
| Μέγιστο μήκος ακολουθίας | 1.024 tokens |
| Ακρίβεια | Bfloat16, χωρίς κβαντοποίηση |
| Απώλεια | Μόνο στα tokens απόκρισης (completion-only) |
| Σύνολο εκπαίδευσης / επικύρωσης | 831 / 92 (στρωματοποιημένος διαχωρισμός) |
| Διάρκεια / βήματα | 299,5s / 156 |
| Τελική απώλεια εκπαίδευσης / επικύρωσης | 0,0269 / 0,0011 |

**Πίνακας 5.2:** Πλήρης διαμόρφωση της προσαρμογής LoRA

## *5.4 Πειραματική Διάταξη*

### Σχεδιασμός

Κύριος στόχος κατά τον πειραματικό σχεδιασμό ήταν η απομόνωση των επιδράσεων, καθώς δεν επιχειρήθηκε να απαντηθεί το ερώτημα μόνο του πόσο καλά αποδίδει το πλήρες σύστημα, αλλά κυριότερο ποια ήταν η συνεισφορά καθενός από τα επιμέρους συστατικά του, δηλαδή το fine-tuning του μοντέλου, η πρόσβαση στο πραγματικό observation και η στρατηγική δρομολόγησης. Για τον σκοπό αυτό, υιοθετήθηκε ένας παραγοντικός σχεδιασμός 2x2, συμπληρωμένος από ένα baseline, ώστε κάθε σύγκριση μεταξύ γειτονικών διαμορφώσεων να μεταβάλλει έναν και μόνο παράγοντα.

### Οι δύο άξονες και το baseline

Οι τέσσερις διαμορφώσεις του σχεδιασμού 2x2 προκύπτουν ως απάντηση ορισμένων ερωτήσεων: Αρχικά, η πρώτη ερώτηση αφορά το «ποιο μοντέλο αποφασίζει;», συνεπώς, αφορά το μοντέλο το Qwen2.5-3B-Instruct χωρίς fine-tuning (zero-shot) έναντι του fine-tuned μέσω LoRA. Ακόμα, πρέπει να διερωτηθεί το «τί βλέπει το μοντέλο ως τρέχον observation;», δηλαδή είτε θα βλέπει ένα σταθερό, κενό περιεχομένου κείμενο στο placeholder είτε το πραγματικό κείμενο της τρέχουσας σελίδας. Στην περίπτωση του placeholder , το μοντέλο αποφασίζει μόνο με βάση τον στόχο, τις προηγούμενες ενέργειες και τα hints του Δέντρου και συνεπώς μέσω της σύγκρισης αυτών των δύο περιπτώσεων μετράμε πόσο βοηθά το ίδιο το observation. Συνδυάζοντας, συνεπώς, τις δύο απαντήσεις ανά δύο, προκύπτουν οι τέσσερις διαμορφώσεις, όπου και στις τέσσερις από αυτές η δρομολόγηση είναι always-decide. Η πέμπτη διαμόρφωση αφορά το baseline, όπου χρησιμοποιεί την ιεραρχική δρομολόγηση με zero-shot μοντέλο και placeholder observation. Διαφέρει, δηλαδή, από την διαμόρφωση zero-shot/placeholder μόνο ως προς την δρομολόγηση και γι αυτόν τον λόγο η σύγκρισή τους απομονώνει καθαρά την επίδραση της στρατηγικής του always-decide που επιλέχθηκε.

### Ταυτόσημο σύνολο ελέγχου – έγκυρες κατά ζεύγη συγκρίσεις

Όλες οι διαμορφώσεις που εξηγήθηκαν παραπάνω, αξιολογούνται στα ταυτόσημα 45 σημεία ελέγχου των 36 trajectories, με το ίδιο seed. Από αυτές τις διαμορφώσεις που εκτελούνται ως χωριστές, πλήρεις αξιολογήσεις, παράγεται ένα αρχείο καταγραφής ανά βήμα (ταυτότητα trajectory, προβλεφθείσα και ορθή ενέργεια, ορθότητα, διαδρομή απόφασης), από το οποίο προκύπτουν όλες οι μετρικές και οι στατιστικές αναλύσεις που ακολουθούν. Αξίζει να σημειωθεί σε αυτό το σημείο, πως η ταύτιση των σημείων ελέγχου δεν είναι κάποια λεπτομέρεια, αλλά είναι αυτή που καθιστά έγκυρες τις κατά

ζεύγη συγκρίσεις που θα αναλυθούν στην επόμενη ενότητα, όπου κάθε διαφορά υπολογίζεται επί των ίδιων ακριβώς προβλημάτων.

### Μετρική βήματος

Η βασική μετρική της αξιολόγησης είναι η ακρίβεια επιπέδου βήματος (step-level), δηλαδή το ποσοστό των σημείων ελέγχου όπου η προβλεφθείσα ενέργεια ταυτίζεται με την ορθή, όπου στη συνέχεια η ταύτιση αυτή κρίνεται από συνάρτηση τριών επιπέδων:

-ακριβής ισότητα συμβολοσειρών (αφού έχει προηγηθεί κανονικοποίηση πεζών, κενών)

-έλεγχος ίδιου ρήματος ενέργειας με σχέση υποσυμβολοσειράς στο όρισμα (ώστε με αυτόν τον τρόπο π.χ. μία ενέργεια επιλογής χρώματος να αναγνωρίζεται ως ορθή όταν το προβλεφθέν όρισμα αποτελεί κανονικοποιημένη μορφή του ορθού)

-ομοιότητα ακολουθιών με threshold 0,8

Η χαλάρωση αυτή κρίνεται αναγκαία λόγω επιφανειακών παραλλαγών στη μορφή των ορισμάτων, ωστόσο εισάγει και μια ελαφρά επιείκεια στην απόλυτη τιμή της μετρικής. Τέλος, κρίσιμο για την εγκυρότητα των συγκρίσεων είναι πως η ίδια ακριβώς συνάρτηση εφαρμόζεται σε όλες τις διαμορφώσεις, ώστε όποια επιείκεια υπάρχει τότε αυτή να κατανέμεται ισότιμα και να μην ευνοεί καμία πλευρά της σύγκρισης.

### Μετρική trajectory

Συμπληρωματικά με την μετρική ανά βήμα, ορίζεται και η ακριβής αντιστοίχιση επιπέδου trajectory, δηλαδή ένα trajectory προσμετράτε ως επιτυχές μόνο εάν όλα τα σημεία ελέγχου του κριθούν ορθά. Με αυτόν τον τρόπο εισάγουμε μια αυστηρότερη μετρική που τιμωρεί τη διασπορά των λαθών, καθώς ένα σύστημα που συγκεντρώνει τα λάθη του σε λίγα trajectories θα υπερτερεί έναντι ενός με ίδιο συνολικό αριθμό λαθών διασκορπισμένων ευρέως, μια διάκριση με ιδιαίτερη πρακτική σημασία, αφού η εκτέλεση μίας και μόνο λανθασμένης ενέργειας αρκεί για να εκτροχιάσει το task.

### Αναγκαίος περιορισμός ερμηνείας

Οι δύο μετρικές που αξιολογήθηκαν προηγουμένως, αξιολογούν κάθε βήμα υπό το αληθές προηγούμενο του καταγεγραμμένου trajectory, δηλαδή το ιστορικό και τα observations κάθε σημείου ελέγχου προέρχονται από επιτυχημένο trajectory, ανεξαρτήτως του τι θα είχε συμβεί εάν το σύστημα είχε πράγματι εκτελέσει τις δικές του προηγούμενες αποφάσεις, χάνεται με λίγα λόγια η δυνατότητα διάδοσης του σφάλματος.

Αυτό, καθιστά τις μετρικές αυτές κατάλληλες για ελεγχόμενη σύγκριση διαμορφώσεων , ωστόσο την καθιστά και μη συγκρίσιμη με μετρικές διαδραστικής επιτυχίας από άκρο σε άκρο (end-to-end), όπως τα ποσοστά επιτυχίας του WebShop, που αναγράφονται στην βιβλιογραφία και όπου εκεί το σφάλμα ενός βήματος μεταβάλλει όλα τα επόμενα. Οι κλίμακες των δύο ειδών μετρικών δεν πρέπει, συνεπώς, να αντιπαραβάλλονται ευθέως.

**Στατιστική μεθοδολογία**

Δεδομένου πως τα 45 σημεία ελέγχου δεν είναι ανεξάρτητα, αλλά ομαδοποιούνται σε 36 trajectories και τα βήματα καθενός μοιράζονται στόχο και suffix, τα διαστήματα εμπιστοσύνης υπολογίζονται με Clustered Bootstrap σε επίπεδο trajectory, όπου σε κάθε επανάληψη επαναδειγματοληπτούνται με επανάθεση τα 36 trajectories (όχι τα μεμονωμένα βήματα), συγκεντρώνονται στη συνέχεια τα βήματα των επιλεγμένων trajectories και υπολογίζεται η μετρική. Συνεπώς, από 10.000 επαναλήψεις προκύπτει το 95% διάστημα εμπιστοσύνης. Για τις διαφορές μεταξύ των διαμορφώσεων, ακόμα, εφαρμόζεται κατά ζεύγη σχήμα, ώστε το διάστημα εμπιστοσύνης της διαφοράς να αντανακλά τη διαφορά επιδόσεων επί κοινών προβλημάτων και όχι τη διακύμανση της επιλογής προβλημάτων. Διάστημα εμπιστοσύνης διαφοράς που αποκλείει το μηδέν ερμηνεύεται ως στατιστικά σημαντική επίδραση στο επίπεδο 5%.

| Διαμόρφωση | Δρομολόγηση | Μοντέλο | Παρατήρηση | Ρόλος |
|---|---|---|---|---|
| Baseline | Ιεραρχική | Zero-Shot | Placeholder | Επίδραση δρομολόγησης |
| Z/P | Always-decide | Zero-Shot | Placeholder | Κοινή αναφορά |
| Z/R | Always-decide | Zero-Shot | Πραγματική (Real) | Επίδραση observation |
| F/P | Always-decide | Fine-tuned | Placeholder | Επίδραση fine-tuning |
| F/R | Always-decide | Fine-tuned | Πραγματική (Real) | Πλήρες σύστημα |

**Πίνακας 5.3:** Οι πέντε πειραματικές διαμορφώσεις

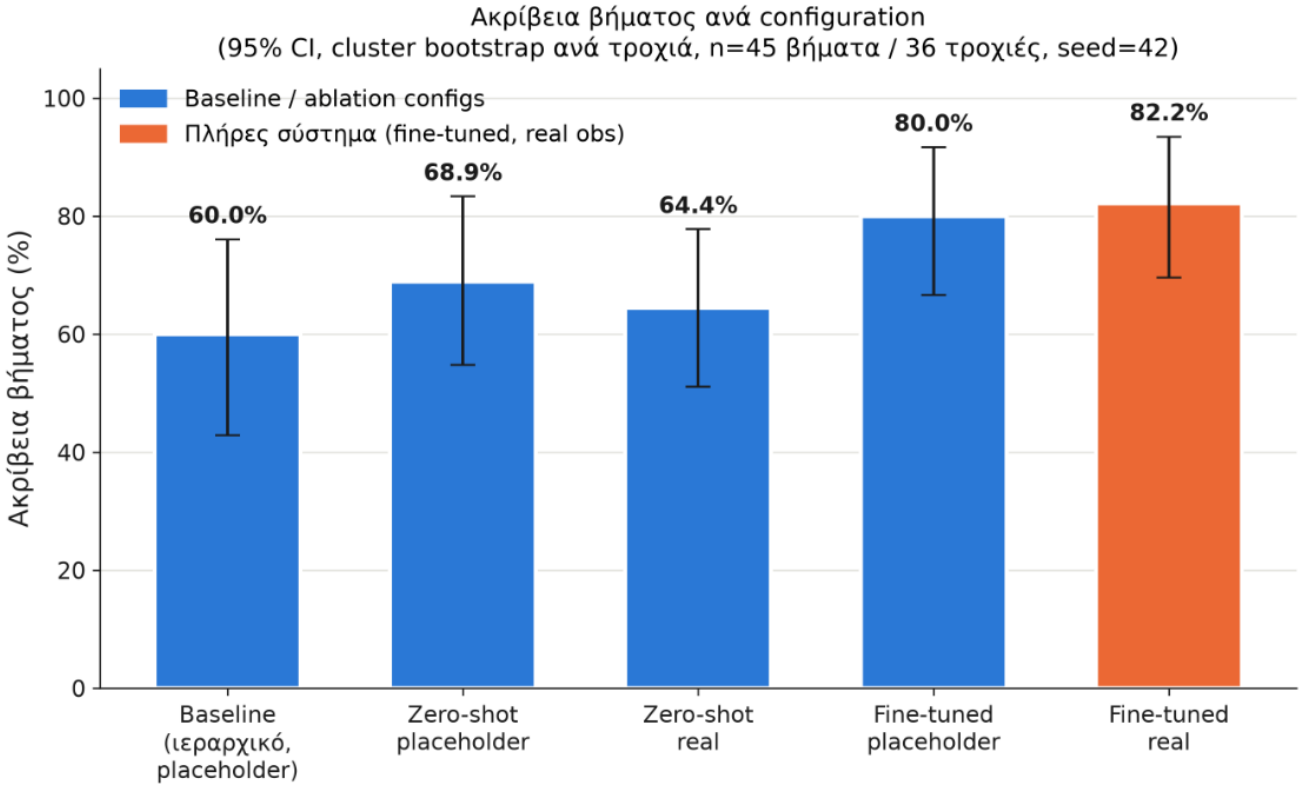
## *5.5 Κύρια Αποτελέσματα*

**Πίνακας αποτελεσμάτων**

Ο Πίνακας 5.4 όπως φαίνεται στη συνέχεια παρουσιάζει την ακρίβεια επιπέδου step και για τις πέντε διαμορφώσεις, επί των ταυτόσημων 45 σημείων ελέγχου. Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα εξής τρία σημεία:

1. Το πλήρες σύστημα (fine-tuned μοντέλο, real observation) επιτυγχάνει την υψηλότερη ακρίβεια, 82,2%.
2. Και οι δύο fine-tuned διαμορφώσεις (80,0% και 82,2%) ξεπερνούν καθαρά όλες τις zero-shot, υποδεικνύοντας τη θετική επίδραση του fine-tuning.
3. Λόγω του περιορισμένου μεγέθους του καθαρού συνόλου αξιολόγησης (n=45), τα διαστήματα εμπιστοσύνης είναι ευρέα, συνεπώς τα ευρήματα που ακολουθούν παρουσιάζονται ως ενδεικτικά (exploratory), με έμφαση στη συνέπεια των επιδράσεων παρά στην τυπική στατιστική σημαντικότητα.

Η Εικόνα 5.2 απεικονίζει τα ίδια αποτελέσματα γραφικά, μαζί με τα 95% διαστήματα εμπιστοσύνης κάθε διαμόρφωσης.




**Εικόνα 5.2:** Ακρίβεια επιπέδου step ανά διαμόρφωση, με 95% διαστήματα εμπιστοσύνης (Cluster Bootstrap ανά trajectory, 10.000 επαναλήψεις)

**Επίδραση της δρομολόγησης (+8,9 μονάδες)**
Η πρώτη σύγκριση αντιπαραθέτει το baseline (ιεραρχική δρομολόγηση, 60,0%) με τη διαμόρφωση που διαφέρει μόνο ως προς τη δρομολόγηση (always-decide, zero-shot placeholder, 68,9%), με την μετάβαση να αποφέρει +8,9 μονάδες (95% CI [−6,7, +25,6]). Παρότι το διάστημα εμπιστοσύνης περιλαμβάνει το μηδέν λόγω του μικρού μεγέθους δείγματος, η θετική κατεύθυνση επιβεβαιώνει ποιοτικά το επιχείρημα της Ενότητας 4.4: όταν το Δέντρο αποφασίζει αυτόνομα, η επίδοση προσεγγίζει το επίπεδο του σκέτου Δέντρου, καθώς οι αυτόνομες αποφάσεις του δεν λαμβάνουν υπόψη την τρέχουσα κατάσταση. Η αξία, δηλαδή, του Δέντρου δεν βρίσκεται στην αυτονομία του αλλά στην συμβουλή του.

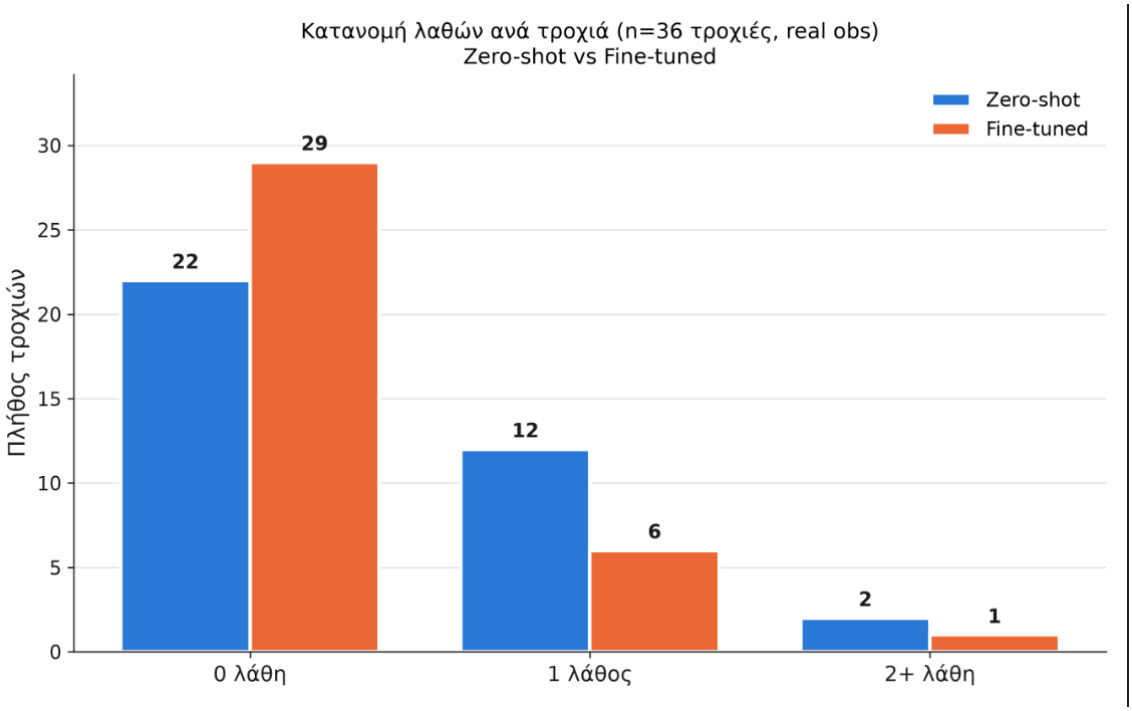
**Επίδραση του fine-tuning**
Η δεύτερη σύγκριση αποτιμά το fine-tuning χωριστά για κάθε είδος observation, δηλαδή με real observation το fine-tuning προσθέτει +17,8 μονάδες (64,4→82,2), με 95% CI [0,0, +35,6], ενώ με placeholder +11,1 μονάδες, με 95% CI [−2,2, +24,5]. Το σημείο εκτίμησης είναι σε αμφότερες τις περιπτώσεις σαφώς θετικό και μεγάλο, ωστόσο τα διαστήματα εμπιστοσύνης (λόγω του μικρού μεγέθους δείγματος) αγγίζουν ή περιλαμβάνουν το μηδέν, οπότε η επίδραση δεν φτάνει το όριο της τυπικής στατιστικής σημαντικότητας στο επίπεδο 5%. Η θετική κατεύθυνση της επίδρασης, εντούτοις, παραμένει συνεπής, όπως τεκμηριώνεται και στη συνέχεια.

**Επίδραση του observation**
Η επίδραση της πρόσβασης στο πλήρες observation αποδεικνύεται μικρή και μη συστηματική στο τρέχον σύνολο: για το μη fine-tuned μοντέλο είναι ελαφρώς αρνητική (−4,4 μονάδες), ενώ για το fine-tuned ελαφρώς θετική (+2,2 μονάδες), με κανένα από τα δύο διαστήματα εμπιστοσύνης να μην αποκλείει το μηδέν. Το εύρημα υποδεικνύει ότι, για ένα μικρό μοντέλο, η επιπλέον πληροφορία της εκτενούς σελίδας δεν μεταφράζεται αυτόματα σε καλύτερη απόφαση, ενδεχομένως λόγω του θορύβου που εισάγει.

**Συνέπεια της επίδρασης**
Αν και καμία μεμονωμένη μέτρηση δεν φτάνει την τυπική σημαντικότητα στο μικρό αυτό σύνολο, η επίδραση του fine-tuning είναι αξιοσημείωτα συνεπής σε πολλαπλές ανεξάρτητες αναλύσεις με +17,8 μονάδες σε επίπεδο βήματος, +19,4 σε επίπεδο trajectory και μείωση των συνολικών σφαλμάτων κατά περίπου 50%. Η σταθερότητα του θετικού προσήμου και του μεγέθους σε διαφορετικές μετρικές αποτελεί ισχυρότερη ένδειξη υπέρ της πραγματικότητας της επίδρασης απ' ό,τι θα υποδήλωνε μια μεμονωμένη οριακή τιμή.

**Αποτελέσματα επιπέδου trajectory**
Η αυστηρότερη μετρική της ακριβούς αντιστοίχισης trajectory συγκρίνει τα δύο σκέλη πραγματικού observation και το αποτέλεσμα είναι εξίσου εντυπωσιακό αφού το ποσοστό των trajectories όπου όλα τα βήματα κρίθηκαν ορθά ανέρχεται από 61,1% (22/36) για το μη fine-tuned μοντέλο σε 80,6% (29/36) για το fine-tuned, βελτίωση +19,4 μονάδων με 95% CI [0,0, +38,9]. Η κατανομή του πλήθους λαθών ανά trajectory (Εικόνα 5.3) αποκαλύπτει τον μηχανισμό της βελτίωσης: τα τέλεια trajectories (χωρίς κανένα λάθος) αυξάνονται από 22 σε 29, ενώ τα trajectories με ένα λάθος μειώνονται από 12 σε 6 και τα trajectories με δύο ή περισσότερα λάθη ήταν ήδη ελάχιστα και παραμένουν χαμηλά (2

→ 1). Το fine-tuning, δηλαδή, μετακινεί trajectories προς τη ζώνη της πλήρους επιτυχίας, μια ιδιότητα με ιδιαίτερη αξία υπό το πρίσμα της διαδραστικής εκτέλεσης, όπου κάθε λάθος απειλεί να εκτροχιάσει ολόκληρο το task. Ένα σύστημα, δηλαδή, που συγκεντρώνει την ορθότητά του σε ακέραια trajectories είναι λειτουργικά προτιμότερο από ένα με ίδιο μέσο όρο αλλά διάσπαρτα λάθη




**Εικόνα 5.3:** Κατανομή των 36 trajectories κατά πλήθος λαθών, πριν και μετά το fine-tuning (real observation)

**Σύνοψη**

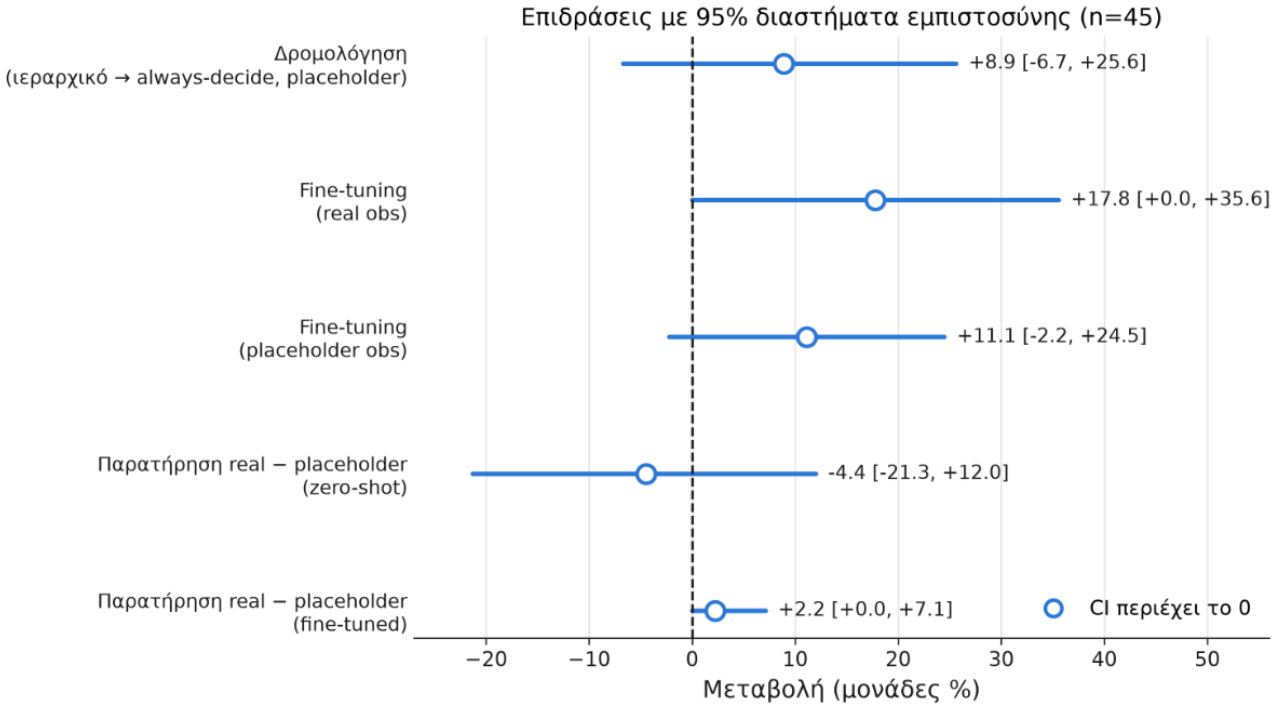
Συνολικά, τα αποτελέσματα υποστηρίζουν, σε επίπεδο ενδείξεων, τους σχεδιαστικούς ισχυρισμούς του Κεφαλαίου 4 πως το fine-tuning βελτιώνει σταθερά την αξιοποίηση των υποδείξεων, η στρατηγική always-decide υπερτερεί της ιεραρχικής, και το πλήρες σύστημα (fine-tuned, real observation) επιτυγχάνει την υψηλότερη επίδοση (82,2%). Ωστόσο, λόγω του περιορισμένου μεγέθους του καθαρού συνόλου αξιολόγησης, οι επιδράσεις δεν φτάνουν την τυπική στατιστική σημαντικότητα και ο περιορισμός αυτός και οι τρόποι άρσης του συζητούνται στην Ενότητα 6.2. Η συμπεριφορά του συστήματος σε καταστάσεις όπου η σωστή ενέργεια λείπει από τα hints εξετάζεται στην Ενότητα 5.8 και αποκαλύπτει σημαντικό περιορισμό του fine-tuning.

| Διαμόρφωση | Ακρίβεια | 95% CI |
|---|---|---|
| Baseline (Ιεραρχική, zero-shot, placeholder) | 60,0% | [42,9, 76,1] |
| Zero-shot, placeholder | 68,9% | [54,8, 83,3] |
| Zero-shot, real observation | 64,4% | [51,1, 77,8] |
| Fine-tuned, placeholder | 80,0% | [66,7, 91,7] |
| Fine-tuned, real observation (πλήρες σύστημα) | 82,2% | [69,6, 93,5] |

**Πίνακας 5.4:** Ακρίβεια επιπέδου step ανά διαμόρφωση (n=45 σημεία, 36 trajectories, seed=42)

| Επίδραση | Σύγκριση | Δ | 95% CI | Σημαντική; |
|---|---|---|---|---|
| Δρομολόγηση | Baseline → Z/P | +8,9 | [−6,7, +25,6] | όχι |
| Fine-tuning (placeholder) | Z/P → F/P | +11,1 | [−2,2, +24,5] | όχι |
| Fine-tuning (real observations) | Z/R → F/R | +17,8 | [0,0, +35,6] | Οριακά όχι |
| Trajectory (real observations) | Z/R → F/R | +19,4 | [0,0, +38,9] | Οριακά όχι |

**Πίνακας 5.5:** Οι τέσσερις επιδράσεις με τα κατά ζεύγη διαστήματα εμπιστοσύνης (τα διαστήματα εμπιστοσύνης είναι ευρέα λόγω του μεγέυθους δειγμάτων(n=45))




**Εικόνα 5.4:** Οι τέσσερις επιδράσεις με τα κατά ζεύγη 95% διαστήματα εμπιστοσύνης. Τα ευρέα διαστήματα αντανακλούν το μικρό μέγεθος δείγματος (n=45)· καμία επίδραση δεν αποκλείει τυπικά το μηδέν.

## *5.6 Ανάλυση Σφαλμάτων*

**Σκοπός και μέθοδος**
Η ανάλυση των σφαλμάτων είναι μια διαδικασία ιδιαίτερα χρήσιμη, καθώς απαντά στο ποια είδη λαθών κάνει το σύστημα, ποια εξ αυτών διορθώνει το fine-tuning και ποια επιμένουν. Για τον σκοπό αυτό, τα λάθη των δύο διαμορφώσεων με real observations (16 λάθη του zero-shot μοντέλου και 8 του fine-tuned, επί των ίδιων 45 σημείων) ταξινομούνται αυτόματα σε κατηγορίες, με ιεραρχικό (μη επικαλυπτόμενο) έλεγχο, δηλαδή κάθε λάθος αποδίδεται στην πρώτη κατηγορία της οποίας τα κριτήρια ικανοποιεί.

**Κατηγορίες**
Οι κατηγορίες αυτές διακρίνονται στις εξής:

**- parsing_format:** όταν η απόκριση του μοντέλου δεν περιέχει καμία έγκυρη συντακτικά ενέργεια

**- wrong_action_type:** όταν το ρήμα της ενέργειας διαφέρει από το ορθό (π.χ. search αντί για click)

**- wrong_id_or_sku:** όταν το όρισμα αφορά κωδικό προϊόντος και είναι εσφαλμένο

**- wrong_navigation:** όταν το όρισμα αφορά στοιχείο πλοήγησης (buy now, back to search, next κ.ο.κ.)

**- wrong_attribute:** όταν το όρισμα αφορά χαρακτηριστικό προιοντος (π.χ. χρώμα, μέγεθος, τιμή κτλ.)

**- other:** για οτιδήποτε δεν εμπίπτει στα ανωτέρω

Συμπληρωματικά, σε κάθε λάθος αποδίδεται ανεξάρτητα η εγκάρσια σήμανση goal-blind, δηλαδή η περίπτωση όπου το ορθό όρισμα εμφανίζεται αυτολεξεί μέσα στο κείμενο του στόχου, ενώ το προβλεφθέν όχι και συνεπώς το μοντέλο είχε την απάντηση έτοιμη, στη διατύπωση του ίδιου του task, αλλά δεν την αξιοποίησε.

**Αποτελέσματα**
Ο Πίνακας 5.6 και η Εικόνα 5.5 παρουσιάζουν την κατανομή. Στο zero-shot μοντέλο κυριαρχούν τα λάθη κωδικών προϊόντος (wrong_id_or_sku, 10 λάθη) και ακολουθούν τα λάθη χαρακτηριστικών (wrong_attribute, 6), ενώ δεν παρατηρήθηκαν λάθη πλοήγησης ή μορφής στο συγκεκριμένο σύνολο. Μετά το fine-tuning, η συνολική μείωση των λαθών (16 → 8) προέρχεται σχεδόν εξ ολοκλήρου από τα λάθη κωδικών, τα οποία σχεδόν εξαλείφονται (10 → 1). Αντιθέτως, τα λάθη χαρακτηριστικών δεν μειώνονται (6 → 7) και αναδεικνύονται στην κύρια εναπομένουσα πρόκληση, συνιστώντας 7 από τα 8 λάθη του πλήρους συστήματος. Το μοτίβο αυτό υποδεικνύει ότι το fine-tuning διορθώνει αποτελεσματικά τα λάθη που επιδέχονται εκμάθηση κανόνα, όπως η ορθή αντιγραφή ενός κωδικού προϊόντος, ενώ η λεπτή διάκριση χαρακτηριστικών παραμένει δύσκολη.

**Επιμονή των λαθών χαρακτηριστικών**

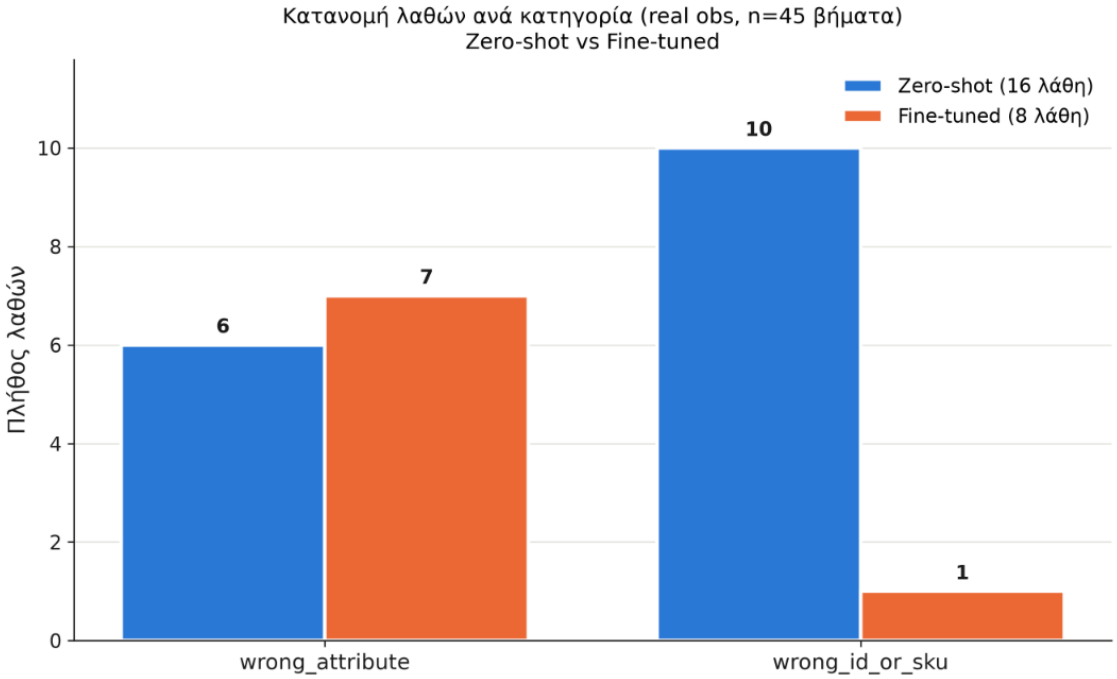
Το εύρημα αυτό δεν προκαλεί έκπληξη υπό το πρίσμα της Ενότητας 4.2, όπου οι πολυπληθέστεροι κόμβοι του δέντρου αντιστοιχούν ακριβώς σε σελίδες επιλογής χαρακτηριστικών, όπου δεκάδες μορφολογικά παρόμοιες ενέργειες (π.χ. click[black], click[navy blue] κ.ο.κ.) ανταγωνίζονται στο ίδιο σημείο απόφασης (ο κορυφαίος κόμβος υπενθυμίζουμε πως καταγράφει 13 διακριτές ενέργειες). Εκεί, ούτε οι ιστορικές υποδείξεις ξεχωρίζουν εύκολα τη σωστή επιλογή (οι κατανομές είναι διάχυτες σε πολλά παρόμοια ορίσματα), ούτε η μορφή της ενέργειας προσφέρει κανόνα προς εκμάθηση, καθώς η ορθή απόφαση απαιτεί λεπτή αντιστοίχιση της εκφώνησης του στόχου με τις διαθέσιμες επιλογές σελίδας. Χαρακτηριστικό παράδειγμα αποτέλεσε ήδη το βήμα της Ενότητας 4.4 όπου το zero-shot μοντέλο επέλεξε click[navy blue] αντί της ορθής επανεπιλογής click[3x-large]. Αξιοσημείωτο είναι, μάλιστα, ότι το fine-tuned μοντέλο εμφανίζει αυξημένη τάση σφαλμάτων τύπου goal-blind, αφού 6 από τα 7 εναπομένοντα λάθη χαρακτηριστικών αφορούν όρισμα που υπήρχε αυτολεξεί στην εκφώνηση του στόχου, χωρίς ωστόσο να επιλεγεί. Το γεγονός αυτό υποδηλώνει ότι το fine-tuning, ενισχύοντας την εμπιστοσύνη του μοντέλου στις ιστορικές υποδείξεις, ενδέχεται να το καθιστά λιγότερο προσεκτικό ως προς την ίδια τη διατύπωση του στόχου, μια ισορροπία που δεδομένου και του μικρού αριθμού των σχετικών λαθών, αποτελεί ενδιαφέρον σημείο για περαιτέρω διερεύνηση.

**Σχεδόν αυστηρή βελτίωση**

Η σύγκριση των δύο συνόλων λαθών σε επίπεδο μεμονωμένου βήματος αποκαλύπτει μια ιδιότητα ισχυρότερη από τη συνολική μείωση: το fine-tuning διόρθωσε 12 βήματα που το zero-shot μοντέλο έκρινε εσφαλμένα, ενώ «χάλασε» 4 βήματα που το zero-shot απαντούσε ορθά, καθαρή βελτίωση 8 βημάτων. Η βελτίωση είναι επομένως ουσιαστική αν και όχι απόλυτη, αφού η πλειονότητα των διορθώσεων αφορά λάθη κωδικών, ενώ οι λίγες νέες αστοχίες εντοπίζονται κυρίως στη διάκριση χαρακτηριστικών.

| **Κατηγορία** | **Zero-shot (16)** | **Fine-tuned (8)** | **Goal-blind (fine-tuned)** |
|---|---|---|---|
| wrong_attribute | 6 | 7 | 6/7 |
| wrong_id_or_sku | 10 | 1 | 0/1 |
| wrong_navigation | 0 | 0 | -- |
| parsing_format | 0 | 0 | -- |
| **Σύνολο** | 16 | 8 | -- |

**Πίνακας 5.6:** Κατανομή λαθών ανά κατηγορία για τις δύο διαμορφώσεις με real observations




**Εικόνα 5.5:** Κατανομή Λαθών ανά Κατηγορία, πριν και μετά το fine-tuning (real observations)

## *5.7 Συμπληρωματικά ευρήματα*

**Σκοπός**

Τα κύρια αποτελέσματα της παρούσας εργασίας όπως αναλύθηκαν παραπάνω τεκμηρίωσαν πως το υβριδικό σύστημα λειτουργεί ορθά και τα ευρήματα της παρούσας ενότητας φωτίζουν τους επιμέρους ρόλους των δύο συστατικών και τους λόγους για τους οποίους το fine-tuning δεν μπορούσε να αντικατασταθεί από απλούστερα μέσα.

**Ο ρόλος των hints ως γείωση**

Ο ρόλος των hints του Δέντρου είναι θεμελιώδης για τη γείωση (grounding) του μοντέλου στο πεδίο των έγκυρων ενεργειών, καθώς χωρίς αυτές, το μοντέλο στερείται πηγής για ορίσματα που αδυνατεί να παραγάγει από μόνο του, τυπικά κωδικούς προϊόντων και συγκεκριμένες ονομασίες επιλογών, αφού δεν υπάρχει πηγή από την οποία να τα αντλήσει. Στο πλήρες σύστημα, τα hints είναι ακριβώς αυτά που μεταφέρουν το λεξιλόγιο των έγκυρων ενεργειών από την ιστορική εμπειρία στο σημείο της απόφασης, συνεπάγοντας πως το Δέντρο δεν προσφέρει απλώς στατιστική καθοδήγηση, αλλά γείωση (grounding) στο πεδίο δράσης, λειτουργία που το γλωσσικό μοντέλο δεν μπορεί να επιτελέσει αυτόνομα.

Η συνεισφορά αυτή μετρήθηκε και άμεσα, σε διερευνητική αξιολόγηση εκτός του αρχικού πειραματικού σχεδιασμού: το μη fine-tuned μοντέλο με real observation αξιολογήθηκε στα 113 σημεία χωρίς καμία υπόδειξη. Η ακρίβεια έπεσε από 78,8% σε

21,2% (διαφορά +57,5 μονάδες υπέρ των υποδείξεων, 95% CI [+49,1, +66,4]), ενώ καμία trajectory δεν ολοκληρώθηκε χωρίς λάθος (0,0% έναντι 68,1%). Η πτώση οφείλεται κυρίως στην αδυναμία παραγωγής έγκυρης ενέργειας: 73 από τις 113 αποκρίσεις δεν περιείχαν συντακτικά έγκυρη ενέργεια, ενώ ανάμεσα στις έγκυρες η ακρίβεια ήταν 60,0%. Το ίδιο αποτέλεσμα προέκυψε όταν διατηρήθηκε ακριβώς η ίδια μορφή προτροπής με κενή λίστα υποδείξεων (21,2%· διαφορά από την προτροπή χωρίς υποδείξεις 0,0 μονάδες, 95% CI [−4,9, +4,7]), άρα η πτώση οφείλεται στο περιεχόμενο των υποδείξεων και όχι στη μορφή της προτροπής.

**Το Δέντρο χωρίς το LLM**

Αντιστρόφως, το Δέντρο ως αυτόνομος λήπτης αποφάσεων, μέσω της ιεραρχικής δρομολόγησης, κινείται στη ζώνη του 60% (baseline). Το επίπεδο αυτό δεν είναι ευκαταφρόνητο για μια δομή που κατασκευάζεται σε δευτερόλεπτα, δεν απαιτεί κάρτα γραφικών και αποφασίζει ακαριαία, ωστόσο αποτελεί άνω φράγμα, καθώς η αμιγώς στατιστική ανάκληση του παρελθόντος είναι τυφλή απέναντι στον στόχο και την τρέχουσα κατάσταση. Παράλληλα, ο συνδυασμός των δύο συστατικών, με το Δέντρο να γειώνει και το μοντέλο να κρίνει, αποδίδει την υψηλότερη επίδοση (82,2% με την αρχική εκπαίδευση και 91,1% με την εκπαίδευση της Ενότητας 5.9), συνθέτοντας το εμπειρικό επιχείρημα της συμπληρωματικότητας πως κάθε συνιστώσα προσφέρει κάτι που η άλλη δεν διαθέτει, χωρίς το ένα να αντικαθιστά το άλλο.

**Prompt Engineering**

Πριν από την προσφυγή στο fine-tuning εξετάστηκε επίσης, στη διερευνητική φάση της εργασίας, αν η αδυναμία του μη fine-tuned μοντέλου να αξιοποιεί τις υποδείξεις θεραπεύεται με απλούστερα μέσα, μέσω τριών παραλλαγών σε επίπεδο προτροπής και μηχανισμού υποδείξεων. Καμία δεν απέφερε βελτίωση, ενώ η επιβολή φίλτρου ελάχιστων επισκέψεων επιδείνωσε την ακρίβεια, γεγονός αναμενόμενο αφού σε ένα τόσο αραιό Δέντρο η απόρριψη κόμβων με λίγες μαρτυρίες ισοδυναμεί συχνά με απόρριψη κάθε υπόδειξης. Οι δοκιμές αυτές έγιναν με διαφορετικό γλωσσικό μοντέλο (Llama 3.2 3B), πριν από την καθιέρωση του πρωτοκόλλου σταθερού seed και με προγενέστερη εκδοχή του κώδικα αξιολόγησης, πριν από τη διόρθωση της διαρροής του baseline (Ενότητα 5.2). Για τον λόγο αυτό δεν παρατίθενται αριθμητικά αποτελέσματα, και οι δοκιμές αναφέρονται μόνο ως τεκμηρίωση της σχεδιαστικής πορείας. Η αναγκαιότητα του fine-tuning τεκμηριώνεται από τις ελεγχόμενες συγκρίσεις των Ενοτήτων 5.5 και 5.9.

**Σύνοψη των συμπληρωματικών ρόλων**

Τα συμπληρωματικά ευρήματα ολοκληρώνουν την εικόνα των κύριων αποτελεσμάτων. Το πλήρες σύστημα λειτουργεί με τρόπο ώστε κάθε συστατικό να καλύπτει μια θεμελιώδη αδυναμία του άλλου, το μοντέλο χωρίς το Δέντρο, δηλαδή, δεν διαθέτει το λεξιλόγιο του πεδίου ενώ το Δέντρο χωρίς το LLM δεν διαθέτει την αντίληψη του παρόντος, ενώ το μοντέλο χωρίς fine-tuning δεν έχει μάθει να τα συνδυάζει. Η υβριδική προσέγγιση της αρχιτεκτονικής και η εκπαίδευσή της, δηλαδή, είναι η γεφύρωση των τριών αυτών κενών.

## *5.8 Αξιολόγηση σε καταστάσεις χωρίς ιστορικό προηγούμενο*

**Σκοπός**
Οι αναλύσεις των προηγούμενων ενοτήτων αφορούν βήματα όπου η σωστή μετάβαση υπάρχει ήδη στο Δέντρο. Η ενότητα αυτή αξιολογεί τα ίδια πέντε configs και στο διορθωμένο, πλήρες σύνολο των 113 σημείων, διακρίνοντας δύο καθεστώτα: το *in-distribution* (τα αρχικά 45 σημεία) και το *OOD* (τα 68 πρόσθετα σημεία, όπου η σωστή ενέργεια δεν περιλαμβάνεται στις υποδείξεις). Και στα δύο, η αξιολόγηση γίνεται με τις ίδιες μετρικές, το ίδιο seed και την ίδια μεθοδολογία bootstrap.

**Αποτελέσματα**
Ο Πίνακας 5.7 παρουσιάζει την ακρίβεια ανά καθεστώς. Στο in-distribution το fine-tuned μοντέλο υπερέχει (82,2% έναντι 64,4% του zero-shot με real observation). Στο OOD το μοτίβο αντιστρέφεται: το fine-tuned πέφτει στο 27,9%, ενώ το zero-shot με real observation φτάνει το 88,2%. Η διαφορά είναι −60,3 μονάδες (95% CI [−72,1, −48,5]) και στο πλήρες σύνολο −29,2 μονάδες (95% CI [−42,1, −17,1]), και στις δύο περιπτώσεις με το διάστημα να αποκλείει το μηδέν. Σε επίπεδο trajectory στο πλήρες σύνολο η μείωση είναι −42,0 μονάδες (68,1% → 26,1%, 95% CI [−56,5, −26,1]). Κάθε OOD σημείο προέρχεται από διαφορετικό επεισόδιο, άρα η bootstrap επαναδειγματοληψία γίνεται σε 68 ανεξάρτητες ομάδες. Ο ίδιος διαχωρισμός με κριτήριο την παρουσία της σωστής ενέργειας στα hints (42 σημεία με, 71 χωρίς) δίνει την ίδια εικόνα: 88,1% έναντι 26,8% για το fine-tuned/real, και 69,0% έναντι 84,5% για το zero-shot/real.

**Ο μηχανισμός**
Ανάλυση των 41 OOD βημάτων όπου το zero-shot/real απαντά ορθά και το fine-tuned/real εσφαλμένα δείχνει ότι το fine-tuned αντιγράφει το κορυφαίο hint: σε 36/41 η απάντησή του ταυτίζεται με αυτό και σε 40/41 με ένα από τα πέντε που εμφανίζονται. Τα hints αυτά είναι αδύναμα (βεβαιότητα 0,1–0,2, από μία μόνο εμφάνιση στο σύνολο εκπαίδευσης). Η σωστή τιμή υπάρχει σε 41/41 περιπτώσεις τόσο στον στόχο όσο και ως κουμπί στη σελίδα· το zero-shot τη βρίσκει με ακριβές ταίριασμα σε 39/41, ενώ μόνο 2/41 απαντήσεις του fine-tuned προέρχονται από τον στόχο. Και τα 41 λάθη ανήκουν στην κατηγορία wrong_attribute. Στο OOD, το fine-tuning δεν διόρθωσε κανένα βήμα και χάλασε 41 (fixed / newly broken: 0/41). Η απόδοση του fine-tuned παραμένει ίδια με placeholder και real observation (27,9%), δηλαδή δεν αξιοποιεί τη σελίδα.

**Ερμηνεία και έκταση**

Το εύρημα συνάδει με την ιδιότητα που περιγράφηκε στην Ενότητα 4.5: το μοντέλο εκπαιδεύτηκε αποκλειστικά σε περιπτώσεις όπου η σωστή ενέργεια ήταν διαθέσιμη στα hints ή δεν υπήρχαν hints. Η ερμηνεία αυτή ελέγχεται πειραματικά, με παρέμβαση, στην Ενότητα 5.9. Το εύρημα περιορίζεται επίσης στο συγκεκριμένο είδος βήματος: και τα 68 OOD σημεία είναι επιλογή attribute σε σελίδα προϊόντος, και 21 από τα 41 αποτελούν μονοσήμια επεισόδια χωρίς ιστορικό ενεργειών. Σε αυτά τα βήματα η σωστή τιμή είναι ρητά διαθέσιμη στον στόχο, γεγονός που εξηγεί εν μέρει και τη μεγάλη επίδοση του zero-shot. Συνεπώς η ακρίβεια του zero-shot στο OOD δεν πρέπει να εκληφθεί ως ένδειξη γενικής ικανότητας πλοήγησης.

| Διαμόρφωση | In-dist (45) | OOD (68) | Πλήρες (113) |
|---|---|---|---|
| Baseline | 60,0% [42,9, 76,1] | 41,2% [29,4, 52,9] | 48,7% [39,4, 57,6] |
| Zero-shot, placeholder | 68,9% [54,0, 84,1] | 73,5% [63,2, 83,8] | 71,7% [63,4, 79,8] |
| Zero-shot, real | 64,4% [51,2, 77,8] | 88,2% [79,4, 95,6] | 78,8% [71,8, 85,7] |
| Fine-tuned, placeholder | 80,0% [66,7, 91,7] | 27,9% [17,6, 38,2] | 48,7% [38,2, 58,6] |
| Fine-tuned, real | 82,2% [69,6, 93,5] | 27,9% [17,6, 38,3] | 49,6% [39,2, 59,1] |

**Πίνακας 5.7**: Ακρίβεια με 95% CI

Σημείωση: οι εκτιμήσεις in-distribution είναι ταυτόσημες με εκείνες του Πίνακα 5.4· τα διαστήματα εμπιστοσύνης της Ενότητας 5.5 προέρχονται από την αξιολόγηση των 45 σημείων, ενώ εκείνα του παρόντος πίνακα από την αξιολόγηση των 113 σημείων, και διαφέρουν ελάχιστα λόγω διαφορετικής σειράς τυχαιότητας.

## *5.9 Έλεγχος της υπόθεσης: εκπαίδευση με υποδείξεις που αποκλείουν το ίδιο task*

**Σκοπός και προκαταχώριση**

Η Ενότητα 5.8 απέδωσε την αποτυχία του fine-tuned μοντέλου στις καταστάσεις OOD στον τρόπο κατασκευής των δεδομένων εκπαίδευσης, δηλαδή το Δέντρο από το οποίο ανακτώνταν οι υποδείξεις περιείχε και το ίδιο το επεισόδιο κάθε παραδείγματος, καθώς και τα άλλα επεισόδια του ίδιου task (96 από τα 209 tasks εκπαίδευσης έχουν δύο επεισόδια). Συνεπώς η σωστή ενέργεια βρισκόταν πάντα μέσα στις υποδείξεις και μάλιστα ως κορυφαία υπόδειξη στο 96,7% των βημάτων. Για να ελεγχθεί η υπόθεση με παρέμβαση, σχεδιάστηκε ένα συμπληρωματικό πείραμα, όπου πριν από την παραγωγή οποιουδήποτε δεδομένου καταχωρίστηκαν γραπτώς η υπόθεση, η παρέμβαση, τα σταθερά στοιχεία και τα κριτήρια επιτυχίας (pre-registration), μαζί με δύο

χρονοσημασμένες διευκρινίσεις, επίσης πριν από την παραγωγή δεδομένων. Το κείμενο και οι αλλαγές του είναι καταγεγραμμένα στο αποθετήριο του κώδικα με κρυπτογραφικό hash, ώστε να επαληθεύεται ότι δεν τροποποιήθηκαν εκ των υστέρων.

**Η παρέμβαση**

Η παρέμβαση αφορά αποκλειστικά τα δεδομένα εκπαίδευσης και περιλαμβάνει τρία στοιχεία. Πρώτον, οι υποδείξεις κάθε παραδείγματος ανακτώνται αφού αποκλειστούν όλες οι εμφανίσεις του Δέντρου που προέρχονται από το ίδιο task, ενώ οι συχνότητες, οι μετρητές επισκέψεων και οι βεβαιότητες υπολογίζονται εκ νέου από τις εμφανίσεις που απομένουν (leave-one-task-out). Δεύτερον, ο μηχανισμός μηδενικής ανταμοιβής δεν εισάγει πλέον την ορθή ενέργεια: όταν κανένα σύνολο δεν οδηγεί στη σωστή απάντηση, το παράδειγμα διατηρεί το πλουσιότερο πραγματικό σύνολο υποδείξεων, με ετικέτα την ορθή ενέργεια. Τρίτον, κάθε κλήση του μοντέλου κατά την παραγωγή των δεδομένων λαμβάνει δικό της seed, ώστε το αποτέλεσμά της να μην εξαρτάται από τη σειρά των υπόλοιπων κλήσεων. Όλα τα υπόλοιπα παρέμειναν σταθερά: το Δέντρο και οι υποδείξεις της αξιολόγησης, τα 113 σημεία ελέγχου, οι ρυθμίσεις LoRA, ο διαχωρισμός εκπαίδευσης/επικύρωσης, οι μετρικές και η μεθοδολογία bootstrap. Πριν από την παρέμβαση επαληθεύτηκε ότι το υπολογιστικό περιβάλλον αναπαράγει τα αποτελέσματα της Ενότητας 5.8 με μηδενικές διαφορές ανά βήμα, και ότι η νέα ανάκτηση συμφωνεί με εξαντλητική αναζήτηση (brute-force) σε όλα τα βήματα εκπαίδευσης.

**Έλεγχος της παρέμβασης**

Ο Πίνακας 5.8 συγκρίνει τα δύο σύνολα προσαρμογής. Το κρίσιμο στοιχείο είναι ότι το νέο σύνολο περιέχει 187 παραδείγματα με υποδείξεις που δεν περιλαμβάνουν τη σωστή ενέργεια, ενώ το αρχικό δεν περιείχε κανένα. Παράλληλα, επειδή πολλές σελίδες εμφανίζονται μόνο στο δικό τους task, τα δύο τρίτα των παραδειγμάτων δεν έχουν πλέον καμία υπόδειξη, κάτι που αντιστοιχεί ακριβώς στην κατάσταση ενός αθέατου task. Η εκπαίδευση έγινε με τις ίδιες ακριβώς ρυθμίσεις (830 παραδείγματα εκπαίδευσης και 93 επικύρωσης, 14,97 εκατομμύρια εκπαιδεύσιμες παράμετροι), διήρκεσε περίπου 5 λεπτά και κατέληξε σε απώλεια επικύρωσης 0,0152.

| Μέγεθος | Αρχικό σύνολο | Σύνολο leave-one-task-out |
|---|---|---|
| Παραδείγματα | 923 | 923 |
| Χωρίς υποδείξεις | 220 | 618 |
| Με υποδείξεις | 703 | 305 |
| – με τη σωστή ενέργεια | 703 | 118 |
| – χωρίς τη σωστή ενέργεια | 0 | 187 |
| Βήματα μηδενικής ανταμοιβής | 165 (17,9%), με εισαγωγή της ορθής ενέργειας | 466 (50,5%), χωρίς εισαγωγή (71 με πραγματικές υποδείξεις) |

**Πίνακας 5.8:** Σύνθεση των δύο συνόλων προσαρμογής

**Αποτελέσματα**

Ο Πίνακας 5.9 προσθέτει τις δύο νέες διαμορφώσεις στον Πίνακα 5.7. Το κύριο, προκαταχωρισμένο αποτέλεσμα υποστηρίζεται: στα 68 σημεία OOD, η ακρίβεια με real observation ανεβαίνει από 27,9% σε 92,6% (+64,7 μονάδες, 95% CI [+52,9, +76,5]). Το δευτερεύον κριτήριο, η διατήρηση τουλάχιστον του μισού αρχικού οφελους στο in-distribution σύνολο, επίσης ικανοποιείται και μάλιστα το όφελος αυξάνεται: το νέο μοντέλο φτάνει 91,1% έναντι 64,4% του μη fine-tuned (+26,7 μονάδες, 95% CI [+11,6, +41,9]· 13 διορθώσεις και 1 νέο λάθος), με διάστημα που πλέον αποκλείει το μηδέν. Στο πλήρες σύνολο των 113 σημείων το νέο μοντέλο είναι η καλύτερη διαμόρφωση: 92,0% έναντι 78,8% του μη fine-tuned (+13,3 μονάδες, 95% CI [+6,6, +20,0]) και 49,6% του αρχικού fine-tuned (+42,5 μονάδες, 95% CI [+33,3, +52,4]). Σε επίπεδο trajectory, οι τέλειες trajectories φτάνουν το 88,4%, έναντι 68,1% και 26,1% αντίστοιχα.

| Διαμόρφωση | In-dist (45) | OOD (68) | Πλήρες (113) |
|---|---|---|---|
| Baseline | 60,0% [42,9, 76,1] | 41,2% [29,4, 52,9] | 48,7% [39,4, 57,6] |
| Zero-shot, placeholder | 68,9% [54,0, 84,1] | 73,5% [63,2, 83,8] | 71,7% [63,4, 79,8] |
| Zero-shot, real | 64,4% [51,2, 77,8] | 88,2% [79,4, 95,6] | 78,8% [71,8, 85,7] |
| Fine-tuned, placeholder | 80,0% [66,7, 91,7] | 27,9% [17,6, 38,2] | 48,7% [38,2, 58,6] |
| Fine-tuned, real | 82,2% [69,6, 93,5] | 27,9% [17,6, 38,3] | 49,6% [39,2, 59,1] |
| Fine-tuned LOTO, placeholder | 53,3% [38,3, 70,0] | 76,5% [66,2, 86,8] | 67,3% [57,1, 77,2] |
| **Fine-tuned LOTO, real** | **91,1% [80,0, 100,0]** | **92,6% [85,3, 98,5]** | **92,0% [86,0, 97,2]** |

**Πίνακας 5.9:** Ακρίβεια με 95% CI, με τις δύο διαμορφώσεις της εκπαίδευσης leave-one-task-out (LOTO)

**Ο μηχανισμός**
Η υπερ-εξάρτηση από τις υποδείξεις εξαλείφεται. Στο OOD σύνολο, το νέο μοντέλο διορθώνει 44 βήματα του αρχικού fine-tuned χωρίς να χαλάσει κανένα, και από τα 5 λάθη του μόνο 1 είναι αντιγραφή της κορυφαίας υπόδειξης, έναντι 43 στα 49 του αρχικού. Επιπλέον, η ακρίβειά του είναι σχεδόν ίδια είτε η σωστή ενέργεια βρίσκεται στις υποδείξεις είτε όχι (92,9% έναντι 91,5%), ενώ για το αρχικό fine-tuned οι αντίστοιχες τιμές ήταν 88,1% και 26,8%. Στο in-distribution σύνολο τα λάθη μειώνονται σε 4 (3 χαρακτηριστικών και 1 κωδικού προϊόντος), κανένα τύπου goal-blind.

Αξιοσημείωτη είναι και η συμπεριφορά με placeholder observation: χωρίς την πραγματική σελίδα, το νέο μοντέλο αποδίδει αισθητά χειρότερα στο in-distribution σύνολο από το αρχικό (53,3% έναντι 80,0%, −26,7 μονάδες, 95% CI [−45,7, −7,0]). Το εύρημα αυτό, που δεν ήταν μέρος της προκαταχώρισης, συνάδει με τον μηχανισμό: το νέο μοντέλο έμαθε να στηρίζει την απόφασή του στον στόχο και στη σελίδα, και όχι αποκλειστικά στις υποδείξεις, άρα η αφαίρεση της σελίδας το επηρεάζει περισσότερο.

**Σταθερότητα ως προς το seed εκπαίδευσης**
Για να ελεγχθεί ότι το αποτέλεσμα δεν οφείλεται σε μία συγκεκριμένη εκπαίδευση, η εκπαίδευση επαναλήφθηκε, σε διερευνητική βάση, με δύο ακόμη seeds στα ίδια δεδομένα (Πίνακας 5.10). Το κύριο αποτέλεσμα υποστηρίζεται και στα τρία (+63,2 έως +66,2 μονάδες), το in-distribution όφελος διατηρείται (12 έως 14 βήματα επιπλέον έναντι του μη fine-tuned ), και η ακρίβεια στο πλήρες σύνολο κυμαίνεται από 92,0% έως 94,7%.

| Seed | In-dist (45) | OOD (68) | Πλήρες (113) | Trajectory (πλήρες) | OOD έναντι αρχικού fine-tuned |
|---|---|---|---|---|---|
| 42 | 91,1% | 92,6% | 92,0% | 88,4% | +64,7 [+52,9, +76,5] |
| 1 | 95,6% | 94,1% | 94,7% | 91,3% | +66,2 [+54,4, +76,5] |
| 2 | 93,3% | 91,2% | 92,0% | 88,4% | +63,2 [+51,5, +75,0] |

**Πίνακας 5.10**: Ακρίβεια του fine-tuned LOTO με real observation σε τρία seeds εκπαίδευσης

**Ερμηνεία και όρια**
Το πείραμα μετατρέπει την ερμηνεία της Ενότητας 5.8 από υπόθεση σε ελεγμένο εύρημα: η υπερ-εξάρτηση οφειλόταν στον τρόπο κατασκευής των δεδομένων εκπαίδευσης και διορθώνεται όταν αυτά περιέχουν και λανθασμένες υποδείξεις. Το αρχικό μοντέλο είχε μάθει να επιλέγει ανάμεσα στις υποδείξεις, αλλά όχι να τις απορρίπτει όλες όταν είναι λανθασμένες· το νέο μαθαίνει και τα δύο και τρία όρια πρέπει να σημειωθούν. Πρώτον, η παρέμβαση άλλαξε ταυτόχρονα την ανάκτηση των υποδείξεων, τον μηχανισμό μηδενικής ανταμοιβής και τον τρόπο ορισμού των seeds, οπότε η επίδραση αποδίδεται στο σύνολο της αλλαγής και όχι σε καθένα στοιχείο χωριστά. Δεύτερον, η παραγωγή των δεδομένων εκτελέστηκε μία φορά. Τρίτον, με μόλις 9 λάθη στα 113 σημεία, οι μετρήσεις πλησιάζουν το ανώτατο όριο, και μικρές διαφορές κοντά σε αυτό δεν μπορούν να διακριθούν· για παράδειγμα, η υπεροχή του νέου μοντέλου έναντι του μη fine-tuned στο OOD σύνολο (+4,4 μονάδες, 95% CI [0,0, +10,3]) δεν είναι διακριτή.

# 6

# *Συμπεράσματα και Μελλοντικές Προεκτάσεις*

## *6.1 Σύνοψη Συνεισφορών*

Στην παρούσα εργασία διερευνήθηκε μια υβριδική αρχιτεκτονική με στόχο την αυτόνομη πλοήγηση agents σε περιβάλλοντα Ιστού, η οποία συνδυάζει ένα Γενικευμένο Δέντρο Επιθημάτων με ένα μικρό fine-tuned γλωσσικό μοντέλο. Έχοντας πλέον παρουσιάσει τον σχεδιασμό (Κεφάλαιο 4), αλλά και τα πειραματικά ευρήματα (Κεφάλαιο 5), η ενότητα αυτή ανακεφαλαιώνει τις συνεισφορές της εργασίας με βάση τα αποτελέσματα που την τεκμηριώνουν.

**Υβριδική αρχιτεκτονική με παρέμβαση στο στάδιο της απόφασης**
Η κεντρική συνεισφορά της έρευνας αυτής είναι μια αρχιτεκτονική που τοποθετεί τη συμβολική γνώση όχι στην επεξεργασία των observations όπως κάνουν προγενέστερες προσεγγίσεις και ειδικότερα το LCoW, αλλά στο ίδιο το στάδιο της απόφασης με το GST να παρέχει ιστορικές υποδείξεις και το γλωσσικό μοντέλο να διατηρεί πάντα τον έλεγχο της τελικής απόφασης. Το πλήρες σύστημα επιτυγχάνει ακρίβεια 82,2% σε επίπεδο βήματος με την αρχική εκπαίδευση και 91,1% με την εκπαίδευση leave-one-task-out (92,0% στο πλήρες σύνολο των 113 σημείων), την υψηλότερη από κάθε διαμόρφωση. Πέραν, ακόμα, της επίδοσης, η αρχιτεκτονική αυτή προσφέρει και ένα επιχείρημα αποδοτικότητας αφού η συμβολική δομή του Δέντρου κατασκευάζεται σε δευτερόλεπτα

και το LLM έχει μόλις 3 δισεκατομμύρια παραμέτρους, εκ των οποίων έγινε fine-tuned λιγότερο από το 0,5%.

**Τεκμηρίωση της αναγκαιότητας του fine-tuning**
Στη συνέχεια, αποδείχθηκε πως η αξιοποίηση των ιστορικών υποδείξεων δεν είναι ζήτημα διατύπωσης της προτροπής (prompt engineering) αλλά μιας δεξιότητας κατά την οποία το LLM θα μάθει πως να σταθμίζει τη γνώση. Το μη fine-tuned μοντέλο αγνοεί συχνά τις υποδείξεις ακόμη και όταν αυτές περιέχουν την ορθή ενέργεια, ενώ διερευνητικές παραλλαγές της προτροπής δεν βελτίωσαν την εικόνα (Ενότητα 5.7). Το κενό γεφυρώνεται μέσω fine-tuning, υπό την προϋπόθεση ότι τα δεδομένα εκπαίδευσης περιέχουν και περιπτώσεις όπου οι υποδείξεις είναι λανθασμένες (Ενότητα 5.9).

**Απομόνωση των επιδράσεων μέσω παραγοντικού σχεδιασμού**
Αντί για μια μεμονωμένη μέτρηση επίδοσης, η εργασία υιοθέτησε σχεδιασμό 2x2 συμπληρωμένο από βασική διαμόρφωση, ώστε να απομονωθεί η συνεισφορά κάθε παράγοντα χωριστά:

- Το fine-tuning του μοντέλου, όπου αυξάνει την απόδοση κατά +11,1 έως +17,8 μονάδες αναλόγως του observation
- Η στρατηγική δρομολόγησης always-decide έναντι της ιεραρχικής, όπου προσδίδει +8,9 μονάδες
- Και τέλος η πρόσβαση στο τρέχον observation, με μικρή επίδραση στην αρχική εκπαίδευση αλλά καθοριστική για το μοντέλο της Ενότητας 5.9

Σημειώνεται ότι, λόγω του μικρού μεγέθους του καθαρού συνόλου αξιολόγησης, οι επιδράσεις αυτές αποτελούν ισχυρές ενδείξεις παρά τυπικά επιβεβαιωμένα αποτελέσματα. Αντίθετα, οι επιδράσεις της εκπαίδευσης leave-one-task-out είναι στατιστικά σημαντικές (Ενότητα 5.9).

**Στατιστική τεκμηρίωση και αξιολόγηση σε επίπεδο trajectory**
Όλες οι αναφερόμενες επιδράσεις συνοδεύονται από διαστήματα εμπιστοσύνης υπολογισμένα με cluster bootstrap σε επίπεδο trajectory. Λόγω του περιορισμένου μεγέθους του καθαρού συνόλου αξιολόγησης, τα διαστήματα είναι ευρέα και οι επιδράσεις της αρχικής εκπαίδευσης δεν φτάνουν την τυπική στατιστική σημαντικότητα, αν και η θετική τους κατεύθυνση παραμένει συνεπής σε όλες τις μετρικές· οι επιδράσεις της εκπαίδευσης leave-one-task-out τη φτάνουν (π.χ. +26,7 μονάδες in-distribution, 95% CI [+11,6, +41,9]). Πέραν της ακρίβειας step, η αυστηρότερη μετρική της ακριβούς αντιστοίχισης trajectory έδειξε βελτίωση +19,4 μονάδων, ενώ η ανάλυση της κατανομής των λαθών αποκάλυψε πως το fine-tuning μετακινεί trajectories προς τη ζώνη της πλήρους επιτυχίας, μια ιδιότητα με ιδιαίτερη αξία για τη διαδραστική εκτέλεση.

**Ανάδειξη των συμπληρωματικών ρόλων των συστατικών**
Τέλος, η εργασία τεκμηρίωσε πως τα δύο συστατικά είναι συμπληρωματικά και όχι ανταγωνιστικά, αφού το γλωσσικό μοντέλο χωρίς hints από το Δέντρο στερείται του

λεξιλογίου των έγκυρων ενεργειών του πεδίου (η ακρίβειά του πέφτει από 78,8% σε 21,2%), ενώ το Δέντρο ως αυτόνομος λήπτης αποφάσεων κινείται στη ζώνη του 60%, διότι είναι τυφλό ως προς τον στόχο και την τρέχουσα κατάσταση. Η αξία της αρχιτεκτονικής, συνεπώς, προκύπτει από τον συνδυασμό των δύο.

**Διασφάλιση εγκυρότητας**
Καθοριστικό μέρος της εργασίας υπήρξε η διασφάλιση της εγκυρότητας των αποτελεσμάτων. Κατά τη διερεύνηση εντοπίστηκαν και αντιμετωπίστηκαν τέσσερα ζητήματα που αλλοίωναν τις αρχικές μετρήσεις: δύο της υλοποίησης (διαρροή στο baseline καιαστοχία στην αντιστοίχιση των υποδείξεων σε επίπεδο ακμής) και δύο εγγενή στα δεδομένα του benchmark (διαρροή εντός συμφραζομένων από την ένωση επαναλαμβανόμενων επεισοδίων, και χρήση μερικώς αποτυχημένων trajectories ως ετικετών αξιολόγησης). Η αντιμετώπισή τους με συνεπή μεθοδολογία, και η επαλήθευση της βασικής δομής μέσω εξαντλητικής σύγκρισης (brute-force oracle), διασφαλίζει την αξιοπιστία των αναφερόμενων αποτελεσμάτων και αναδεικνύει πώς τέτοιες παγίδες αξιολόγησης μπορούν να παραγάγουν φαινομενικά ισχυρά αλλά μη έγκυρα ευρήματα.

**Διάγνωση και διόρθωση της υπερ-εξάρτησης από τις υποδείξεις**
Η αξιολόγηση στο πλήρες σύνολο έδειξε αρχικά ότι η θετική επίδραση του fine-tuning περιοριζόταν σε καταστάσεις όπου η σωστή ενέργεια περιλαμβάνεται στις υποδείξεις: όταν απουσίαζε, το fine-tuned μοντέλο αντέγραφε το κορυφαίο hint και υπολειπόταν κατά 60,3 μονάδες του μη fine-tuned . Η αιτία εντοπίστηκε στην κατασκευή των δεδομένων εκπαίδευσης, όπου η σωστή ενέργεια βρισκόταν πάντα στις υποδείξεις, και ελέγχθηκε με προκαταχωρισμένο πείραμα παρέμβασης. Με υποδείξεις εκπαίδευσης που αποκλείουν το ίδιο task, η υπερ-εξάρτηση εξαλείφεται: η ακρίβεια στις καταστάσεις αυτές ανεβαίνει από 27,9% σε 92,6% (+64,7 μονάδες, 95% CI [+52,9, +76,5]), το όφελος στο in-distribution σύνολο αυξάνεται, και το μοντέλο γίνεται η καλύτερη διαμόρφωση στο πλήρες σύνολο (92,0%), με σταθερά αποτελέσματα σε τρία seeds. Το γενικότερο δίδαγμα είναι ότι, όταν ένα μοντέλο εκπαιδεύεται πάνω σε ανακτημένη πληροφορία, η ανάκτηση κατά την εκπαίδευση πρέπει να μοιάζει με την ανάκτηση κατά τη χρήση.

## *6.2 Περιορισμοί*

Η ερμηνεία των αποτελεσμάτων της εργασίας οφείλει να λαμβάνει υπόψιν μια σειρά περιορισμών, οι οποίοι οριοθετούν την εμβέλεια των συμπερασμάτων και ταυτόχρονα υποδεικνύουν κατευθύνσεις για μελλοντική έρευνα.

**Μέγεθος του συνόλου αξιολόγησης**
Ο πλέον καθοριστικός περιορισμός της μελέτης είναι το μικρό μέγεθος του καθαρού συνόλου αξιολόγησης (45 σημεία ελέγχου από 36 trajectories). Το μέγεθος αυτό, που

προκύπτει από τη σχεδίαση exact-matching και το φιλτράρισμα σε πλήρως επιτυχημένα επεισόδια (Ενότητα 5.2), έχει άμεση συνέπεια στη στατιστική δύναμη, διότι τα διαστήματα εμπιστοσύνης είναι ευρέα και αν και οι επιδράσεις εμφανίζουν σταθερά θετική κατεύθυνση και σημαντικό μέγεθος δεν φτάνουν την τυπική στατιστική σημαντικότητα στο επίπεδο 5%. Τα αποτελέσματα, επομένως, παρουσιάζονται ως ισχυρές ενδείξεις (exploratory) παρά ως οριστικά επιβεβαιωμένα ευρήματα. Η αύξηση του συνόλου αξιολόγησης, μέσω ελαστικότερης αντιστοίχισης observations ή διαδραστικής εκτέλεσης, αποτελεί την κύρια προϋπόθεση για ισχυρότερη στατιστική τεκμηρίωση. Οι διαφορές του πειράματος της Ενότητας 5.9, επειδή είναι μεγαλύτερες, είναι στατιστικά σημαντικές παρά το μικρό μέγεθος. Ωστόσο, με λίγα πλέον λάθη (9 στα 113), οι μετρήσεις πλησιάζουν το ανώτατο όριο, και μικρές διαφορές κοντά σε αυτό δεν μπορούν να διακριθούν.

**Η φύση της μετρικής αξιολόγησης**
Ακόμα, ένας σημαντικός περιορισμός της παρούσας εργασίας αφορά τον τρόπο αξιολόγησης. Όλες οι μετρικές υπολογίζονται υπό το αληθές προηγούμενο κάθε trajectory, όπου σε κάθε βήμα το ιστορικό και το observation προέρχονται από το καταγεγραμμένο επιτυχημένο trajectory, ανεξαρτήτως των προηγούμενων αποφάσεων του ίδιου του συστήματος. Η αξιολόγηση, δηλαδή, δεν εμπεριέχει διάδοση σφάλματος (error propagation), κάτι το οποίο περιέχει ένα live rollout σε end-to-end ανάλυση όπως κάνει το LCoW. Αυτό είναι μεθοδολογικά αναγκαίο για την εγκυρότητα της εσωτερικής σύγκρισης, αλλά συνεπάγεται πως η αναφερόμενη ακρίβεια βήματος δεν αντιστοιχεί σε ποσοστό επιτυχίας διαδραστικής εκτέλεσης end-to-end. Σε πραγματική εκτέλεση, δηλαδή, όπου ένα εσφαλμένο βήμα οδηγεί σε καταστάσεις που ενδέχεται να μην έχουν ιστορικό προηγούμενο, η επίδοση αναμένεται χαμηλότερη. Οι αναφερόμενες τιμές, επομένως, δεν είναι ευθέως συγκρίσιμες με μετρικές επιτυχίας σε end-to-end εκτελέσεις και η άρση αυτού του περιορισμού μέσω πραγματικής εκτέλεσης (live rollout) αποτελεί την πρώτη προτεραιότητα μελλοντικής εργασίας.

**Μεταβλητότητα της εκπαίδευσης**
Τα διαστήματα εμπιστοσύνης αποτυπώνουν τη μεταβλητότητα ως προς το σύνολο ελέγχου, μέσω cluster bootstrap επί των trajectories, και όχι τη μεταβλητότητα της ίδιας της εκπαίδευσης. Η αρχική εκπαίδευση εκτελέστηκε με ένα σταθερό seed. Για την εκπαίδευση leave-one-task-out, η εκπαίδευση επαναλήφθηκε με τρία seeds στα ίδια δεδομένα: η ακρίβεια στο πλήρες σύνολο κυμάνθηκε από 92,0% έως 94,7%, και το κύριο αποτέλεσμα επιβεβαιώθηκε σε όλα (Ενότητα 5.9). Η παραγωγή των δεδομένων εκπαίδευσης, ωστόσο, εκτελέστηκε μία φορά, οπότε η μεταβλητότητά της δεν έχει μετρηθεί.

**Ο μηχανισμός των βημάτων μηδενικής ανταμοιβής**
Κατά την κατασκευή του αρχικού συνόλου προσαρμογής, το 17,9% των βημάτων δεν βρήκε σύνολο υποδείξεων που να οδηγεί το μοντέλο στην ορθή ενέργεια, και αντιμετωπίστηκε με εισαγωγή της ορθής ενέργειας ως συνθετικής υπόδειξης με βεβαιότητα 1,0· σε 144 από τα 165 τέτοια παραδείγματα, αυτή ήταν η μοναδική υπόδειξη. Ο μηχανισμός εφαρμόζεται μόνο στην εκπαίδευση, και η αξιολόγηση χρησιμοποιεί αποκλειστικά πραγματικές υποδείξεις του Δέντρου. Ενίσχυσε όμως το μοτίβο «η σωστή απάντηση βρίσκεται στις υποδείξεις», συμβάλλοντας στην υπερ-

εξάρτηση της Ενότητας 5.8. Στο πείραμα της Ενότητας 5.9 ο μηχανισμός αναθεωρήθηκε ώστε να διατηρεί τις πραγματικές υποδείξεις, χωρίς εισαγωγή της ορθής ενέργειας.

**Ακριβής αντιστοίχιση observations**
Η αντιστοίχιση των observations στο Δέντρο είναι ακριβής, καθώς δύο observations θεωρούνται ίδια μόνο εάν το κείμενό τους ταυτίζεται πλήρως. Η επιλογή αυτή προσδίδει ντετερμινιστικό και επαληθεύσιμο χαρακτήρα στη δομή, έχει όμως δύο συνέπειες πως αφενός σχεδόν πανομοιότυπες σελίδες που διαφέρουν σε επουσιώδη σημεία λαμβάνουν διαφορετικά Ids και δεν αντιστοιχίζονται και αφετέρου όπως τεκμηριώθηκε στην Ενότητα 5.2 trajectories ελέγχου με observations άγνωστα κατά την εκπαίδευση αποκλείονται από την αξιολόγηση. Μια πιο ελαστική σημασιολογική αντιστοίχιση θα αύξανε την κάλυψη του Δέντρου, με αντίτιμο μέρος της αυστηρότητάς του.

**Βεβαιότητα και υποδείξεις**
Το visit_count που τροφοδοτεί τον συντελεστή βεβαιότητας των hints καταμετρά ζεύγη trajectory-suffix και όχι διακριτά trajectories ως εκ τούτου ο συντελεστής βεβαιότητας κορέννυται με λιγότερες ανεξάρτητες μαρτυρίες απ' όσες υποδηλώνει η ονομαστική του τιμή. Οι βαθμοί των hints είναι, συνεπώς, χρήσιμοι για τη σχετική κατάταξη εντός ενός κόμβου, αλλά δεν πρέπει να ερμηνεύονται ως καλά βαθμονομημένες πιθανότητες.

**Μεροληψία επιλογής του συνόλου αξιολόγησης**
Πέραν του μεγέθους του, το σύνολο αξιολόγησης υπόκειται σε μια λεπτή μεροληψία επιλογής (selection bias) που αξίζει ρητής αναφοράς. Λόγω της αντιστοίχισης exact-matching, ένα βήμα διατηρείται στο σύνολο ελέγχου μόνον εφόσον η παρατήρηση στην οποία οδηγεί υπάρχει ήδη στο δέντρο, δηλαδή έχει εμφανιστεί κατά την εκπαίδευση. Καθώς η επόμενη αυτή παρατήρηση είναι το αποτέλεσμα της ορθής ενέργειας, το σύνολο των 45 σημείων τείνει να περιλαμβάνει βήματα για τα οποία η ορθή μετάβαση υπάρχει ήδη στο ιστορικό του δέντρου (στην πράξη, 43 από τα 45). Συνεπώς, η αναφερόμενη ακρίβεια πρέπει να ερμηνεύεται ως επίδοση σε καταστάσεις με ιστορικό προηγούμενο, και όχι ως μέτρο γενίκευσης σε εντελώς άγνωστες καταστάσεις. Σημειώνεται ότι πρόκειται για μεροληψία ως προς την επιλογή των σημείων και όχι για διαρροή εντός της προτροπής: η ορθή ενέργεια ουδέποτε εμφανίζεται στο ιστορικό ενεργειών που δίνεται στο μοντέλο (επαληθεύτηκε για όλα τα σημεία). Η αξιολόγηση σε πλήρες σύνολο παρουσιάζεται στην Ενότητα 5.8.

**Ομοιογένεια και ορισμός του OOD συνόλου**
Το OOD σύνολο (68 σημεία) αποτελείται αποκλειστικά από βήματα επιλογής attribute σε σελίδα προϊόντος, και ο διαχωρισμός in-distribution / OOD ορίζεται εκ των υστέρων από το περιεχόμενο των υποδείξεων. Επιπλέον, 21 από τα 41 σχετικά βήματα είναι μονοσήμια επεισόδια, και 2 από τις 41 σωστές απαντήσεις του zero-shot προέκυψαν από το ασαφές ταίριασμα (Ενότητα5.4). Η εξήγηση που προτάθηκε για το φαινόμενο ελέγχθηκε με παρέμβαση (Ενότητα 5.9). Η παρέμβαση όμως άλλαξε ταυτόχρονα τρία στοιχεία (την ανάκτηση των υποδείξεων, τον μηχανισμό μηδενικής ανταμοιβής και τον τρόπο ορισμού των seeds κατά την παραγωγή δεδομένων), οπότε η επίδραση αποδίδεται στο σύνολο της αλλαγής και όχι σε καθένα στοιχείο χωριστά.

**Επικάλυψη tasks στο σύνολο επικύρωσης**
Ο διαχωρισμός των παραδειγμάτων προσαρμογής σε εκπαίδευση και επικύρωση γίνεται σε επίπεδο παραδείγματος, με αποτέλεσμα ορισμένα tasks να εμφανίζονται και στα δύο υποσύνολα (80 στην αρχική εκπαίδευση, 78 στην εκπαίδευση leave-one-task-out). Αυτό εξηγεί τις πολύ χαμηλές τιμές της απώλειας επικύρωσης και επηρεάζει την επιλογή του checkpoint, όχι όμως τα αποτελέσματα ελέγχου, τα οποία αφορούν αθέατα tasks.

**Εύρος της πειραματικής επικύρωσης**
Τέλος, η επικύρωση περιορίζεται σε ένα περιβάλλον (WebShop) και ένα γλωσσικό μοντέλο (Qwen2.5-3B-Instruct). Επιπλέον, η αποτελεσματικότητα της συμβολικής δομής εδράζεται εν μέρει σε ένα χαρακτηριστικό του συγκεκριμένου περιβάλλοντος πως η ύπαρξη επαναλαμβανόμενων μοτίβων observations, όπως οι σελίδες επιλογής χαρακτηριστικών στις οποίες επιστρέφει ο agent. Σε περιβάλλοντα με μικρότερη επανάληψη καταστάσεων, η συνεισφορά του δέντρου ενδέχεται να είναι περιορισμένη. Η γενίκευση των ευρημάτων σε άλλα περιβάλλοντα και μοντέλα παραμένει, επομένως, ανοιχτό ερώτημα.

## *6.3 Μελλοντικές Προεκτάσεις*

**Αξιολόγηση σε πραγματική εκτέλεση (live rollout)**
Η άμεση προτεραιότητα είναι η αξιολόγηση του συστήματος σε πραγματικό, διαδραστικό περιβάλλον WebShop, όπου κάθε ενέργεια εκτελείται πράγματι και το επόμενο observation προκύπτει από αυτήν. Μια τέτοια αξιολόγηση θα αναιρούσε τον κεντρικό περιορισμό της μετρικής που αναφέρθηκε προηγουμένως, αποδίδοντας ποσοστό επιτυχίας end-to-end, ευθέως συγκρίσιμο με τη βιβλιογραφία. Η υποδομή για το βήμα αυτό έχει ήδη προετοιμαστεί ουσιαστικά, αφού το περιβάλλον έχει στηθεί και τεκμηριωθεί πλήρως, ωστόσο, η υλοποίηση αναστάληκε προσωρινά λόγω εξωτερικού εμποδίου, καθώς οι επίσημες πηγές δεδομένων του WebShop κατέστησαν προσωρινά μη διαθέσιμες. Η αποκατάσταση της πρόσβασης στα δεδομένα καθιστά το live rollout άμεσα εκτελέσιμο και σημειώνεται πως το fine-tuning δεν απαιτεί ανάπτυξη νέου μηχανισμού πλαισίωσης παρατηρήσεων (contextualization module), καθώς το σύστημα λειτουργεί επί του ακατέργαστου κειμένου του περιβάλλοντος. Επιπλέον, η διαδραστική εκτέλεση θα επέτρεπε αξιολόγηση σε όλο το φάσμα των tasks, αίροντας τον περιορισμό του μικρού συνόλου αξιολόγησης που επιβάλλει η ακριβής αντιστοίχιση και ενισχύοντας τη στατιστική δύναμη των συμπερασμάτων.

**Ελαστικότερη αντιστοίχιση observations**
Ο περιορισμός της ακριβούς αντιστοίχισης που σημειώθηκε ανωτέρω υποδεικνύει τη διερεύνηση σημασιολογικής ή προσεγγιστικής αντιστοίχισης observations, ώστε σχεδόν πανομοιότυπες σελίδες να ομαδοποιούνται και να μειωθεί με αυτόν τον τρόπο το ποσοστό των άγνωστων observations. Το εγχείρημα αυτό θα αύξανε την κάλυψη που παρέχει το Δέντρο και κατά συνέπεια τη διαθεσιμότητα των hints, με το τίμημα ενός μέρους της ντετερμινιστικής αυστηρότητας που χαρακτηρίζει την τρέχουσα προσέγγιση.

**Στοχευμένη αντιμετώπιση των λαθών χαρακτηριστικών**
Η ανάλυση σφαλμάτων ανέδειξε πως μετά το fine-tuning η κύρια εναπομένουσα πρόκληση είναι η λεπτή διάκριση των χαρακτηριστικών (χρώματα, μεγέθη, τιμές κτλ.) σε σελίδες με μορφολογικά παρόμοιες επιλογές. Μια μελλοντική εργασία θα μπορούσε να εστιάσει ειδικά σε αυτό, στον εμπλουτισμό δηλαδή των υποδείξεων ώστε να αποτυπώνουν καλύτερα τη διάκριση των χαρακτηριστικών, με μηχανισμούς ρητής αντιστοίχισης της εκφώνησης του στόχου στις διαθέσιμες επιλογές της σελίδας ή με στοχευμένη ενίσχυση των παραδειγμάτων εκπαίδευσης αυτής της κατηγορίας. Η κατεύθυνση αυτή στοχεύει ακριβώς στα λάθη χαρακτηριστικών, που συνιστούν τη μεγάλη πλειονότητα (7 από τα 8 στην αρχική εκπαίδευση, 3 από τα 4 στην εκπαίδευση leave-one-task-out) των εναπομεινάντων λαθών του πλήρους συστήματος.

**Ενίσχυση της στατιστικής και πειραματικής βάσης**
Δύο επεκτάσεις που θα ενίσχυαν την αυστηρότητα των αποτελεσμάτων είναι αρχικά η επέκταση των πολλαπλών seeds, που εφαρμόστηκαν ήδη στην εκπαίδευση leave-one-task-out, και στην παραγωγή των δεδομένων εκπαίδευσης, αλλά και μια συστηματική διερεύνηση των υπερπαραμέτρων του fine-tuning, αφού θα αποσαφηνιζόταν η σχέση μεταξύ χωρητικότητας του fine-tuning και της επίδοσης. Χρήσιμη θα ήταν, επίσης, η σύγκριση του Δέντρου Επιθημάτων με απλούστερους μηχανισμούς ανάκτησης ιστορικής εμπειρίας, ώστε να απομονωθεί η ειδική συνεισφορά της συγκεκριμένης δομής έναντι της γενικής ιδέας της ανάκτησης.

**Αποδόμηση της παρέμβασης leave-one-task-out**
Η παρέμβαση της Ενότητας 5.9 άλλαξε ταυτόχρονα την ανάκτηση των υποδείξεων και τον μηχανισμό μηδενικής ανταμοιβής. Πειράματα που αλλάζουν κάθε στοιχείο χωριστά θα έδειχναν πόσο συνεισφέρει το καθένα, ενώ επαναλήψεις της παραγωγής των δεδομένων θα αποτύπωναν τη μεταβλητότητά της. Ενδιαφέρον έχει επίσης η σύνθεση του νέου συνόλου προσαρμογής, στο οποίο τα δύο τρίτα των παραδειγμάτων δεν έχουν υποδείξεις, και το πώς αυτή επηρεάζει την ισορροπία ανάμεσα στην αξιοποίηση και την απόρριψη των υποδείξεων.

**Γενίκευση σε άλλα περιβάλλοντα**
Τέλος, η εφαρμογή της αρχιτεκτονικής σε περιβάλλοντα πέραν του WebShop θα έλεγχε την εγκυρότητα των ευρημάτων και ειδικότερα της εξάρτησης της συμβολικής δομής από την ύπαρξη επαναλαμβανόμενων μοτίβων καταστάσεων, καθώς περιβάλλοντα με πλουσιότερους χώρους ενεργειών και μεγαλύτερη ποικιλία καταστάσεων, όπως πιο σύνθετες εργασίες πλοήγησης στον Ιστό, θα αποτελούσαν κατάλληλο επόμενο πεδίο δοκιμής.

**Επίλογος**
Εν κατακλείδι, η εργασία αυτή κατέδειξε πως μια ρητή συμβολική δομή μπορεί να ενισχύσει ουσιωδώς ένα μικρό γλωσσικό μοντέλο στην πλοήγηση σε περιβάλλοντα Ιστού, όταν η παρέμβαση τοποθετείται στο στάδιο της απόφασης και το μοντέλο εκπαιδεύεται να αξιοποιεί την ιστορική καθοδήγηση. Πέραν της επίδοσης του συγκεκριμένου συστήματος, το ευρύτερο συμπέρασμα είναι πως η συμβολική και η νευρωνική προσέγγιση δεν αποτελούν ανταγωνιστικές εναλλακτικές αλλά συμπληρωματικά συστατικά ,αφού η μία προσφέρει ντετερμινιστική, επιθεωρήσιμη

γείωση και η άλλη ευελιξία και κατανόηση του παρόντος. Παράλληλα, ανέδειξε ότι ο τρόπος κατασκευής των δεδομένων εκπαίδευσης καθορίζει αν το μοντέλο μαθαίνει να κρίνει τις υποδείξεις ή απλώς να τις αντιγράφει, και ότι μια προσεκτική, προκαταχωρισμένη διόρθωση μπορεί να μετατρέψει έναν σημαντικό περιορισμό σε ισχυρό αποτέλεσμα. Η γεφύρωσή τους, μαζί με την εκπαίδευση που την καθιστά λειτουργική, υπήρξε το κεντρικό αντικείμενο της παρούσας εργασίας.

# Βιβλιογραφία- Αναφορές


**Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003).** A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.

**Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020).** Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

**Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020).** Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

**Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022).** Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

**Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017).** Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

**Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022a).** Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

**Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022b).** Emergent abilities of large language models. *Transactions on Machine Learning Research*.

**Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021).** Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

**Lee, D., Lee, J., Kim, K., Tack, J., Shin, J., Teh, Y. W., & Lee, K. (2025, May).** Learning to contextualize web pages for enhanced decision making by LLM agents. In *International Conference on Learning Representations* (Vol. 2025, pp. 81039-81072).

**Weiner, P. (1973).** Linear pattern matching algorithms.

**Liu, E. Z., Guu, K., Pasupat, P., Shi, T., & Liang, P. (2018).** Reinforcement learning on web interfaces using workflow-guided exploration. *arXiv preprint arXiv:1802.08802*.

**Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., ... & Su, Y. (2023).** Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, *36*, 28091-28114.

**Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., ... & Neubig, G. (2024, May).** Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations* (Vol. 2024, pp. 15585-15606).

**Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021).** Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

**Kagaya, T., Yuan, T. J., Lou, Y., Karlekar, J., Pranata, S., Kinose, A., ... & You, Y. (2024).** Rap: Retrieval-augmented planning with contextual memory for multimodal llm agents. *arXiv preprint arXiv:2402.03610*.

**Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October).** Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1-22).

**Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., ... & Lin, J. (2024).** Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

**Shi, T., Karpathy, A., Fan, L., Hernández, J.Z., & Liang, P. (2017).** World of Bits: An Open-Domain Platform for Web-Based Agents. *International Conference on Machine Learning*.

**Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023).** Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, *36*, 8634-8652.

**Ukkonen, E. On-line construction of suffix trees. *Algorithmica* 14, 249–260 (1995).**

**Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017).** Attention is all you need. *Advances in neural information processing systems*, *30*.

**Wang, Z. Z., Mao, J., Fried, D., & Neubig, G. (2024).** Agent workflow memory. *arXiv preprint arXiv:2409.07429*.

**Yao, S., Chen, H., Yang, J., & Narasimhan, K. (2022).** Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, *35*, 20744-20757.

**Zheng, K., Zhou, K., Gu, J., Fan, Y., Wang, J., Di, Z., ... & Wang, X. E. (2022).** Jarvis: A neuro-symbolic commonsense reasoning framework for conversational embodied agents. *arXiv preprint arXiv:2208.13266*.

**Zheng, L., Wang, R., Wang, X., & An, B. (2024, May).** Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *International Conference on Learning Representations* (Vol. 2024, pp. 19036-19066).

**Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., ... & Neubig, G. (2024, May).** Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations* (Vol. 2024, pp. 15585-15606).