Reviewer K9V3

Concern 1: Cross-lingual Model Validation *"How can we know that the core model is cross lingual? If I understand data for all languages have been used when training the model."*

Revision: Added Section 4.3 "Cross-lingual Transfer Validation" with leave-one-language-out experiments (Table 1, lines 330-354). Core model trained on three languages, evaluated on fourth using only adapter training. Results show consistent 3-6% degradation, validating genuine cross-lingual transfer.

Concern 2: Universal Form Analyzer Integration *"The character model is somehow used in the 'Universal Form Analyser' but how? Without this it is impossible to reproduce the experiment."*

Revision: Enhanced Section 3.4 with detailed pipeline description (lines 251-276). Added mathematical formulation (Equation 4) showing integration: *field_type = classify(Adapter_l(f_model(context)))*. Comprehensive implementation details in Appendix E.

Concern 3: Character Prediction Accuracy *"The character-level model shows modest accuracy on masked character prediction."*

Revision: Enhanced discussion in Section 5.1 (lines 401-413) contextualizing that character-level prediction is inherently challenging due to larger candidate space. Demonstrated sufficiency for downstream tasks through comprehensive evaluation.

Reviewer 75Vx

Concern 1: Character-level Convolutional Embeddings *"Provide a more advanced explanation of character-level convolutional embeddings to improve understanding and reproducibility."*

Revision: Added Section 4.1 with detailed mathematical formulation (Equation 7, lines 272-285) and Appendix B.1 (lines 820-842) with PyTorch implementation details, padding strategies, and computational efficiency considerations.

Concern 2: 70/30 Masking Ratio Justification *"The authors chose a 70/30 ratio... but did not explain how this ratio was determined."*

Revision: Added mathematical formulation (Equation 8, lines 293-307) and empirical justification in Section 4.2 (lines 301-307). Tested multiple ratios (50/50, 80/20) before settling on 70/30 based on empirical results across all languages.

Concern 3: Language Adapter Architecture Details *"It is unclear whether and how the method determines which adapter to use and when."*

Revision: Enhanced Section 3.3 and added detailed language detection methodology in Appendix E.1 (lines 1003-1013). Character pattern matching: German (ü,ö,ä,ß), French (accents), Arabic (Unicode ranges), English (default).

Concern 4: Missing Navigation Results *"Section 5.3, four key tasks were identified... However, Chapter 6 (Results) does not include Website Navigation."*

Revision: Added Section 5.6 "Website Navigation Performance" with Table 4 (lines 463-474) showing breakdown: Simple navigation (95%), Cross-domain (85.1-87.6%), addressing abstract claims of 88-95% success rate.

Concern 5: Baseline Comparison Methodology *"The rebuttal does not clarify whether the evaluation involving mBERT was a simple substitution... or if further architectural or pipeline modifications were made."*

Revision: Added Appendix D "Baseline Implementation Details" (lines 950-975) with explicit methodology: identical training data, hyperparameters, and evaluation protocols. No architectural modifications to mBERT, ensuring fair comparison.

Reviewer ivND

Concern 1: Model Selector Implementation Gap *"Line 542 explains that 'model selector is currently heuristic-based', but what does this exactly mean? Was the language model manually chosen from the two options?"*

Revision: Implemented learned model selector throughout:

- Section 3.2 (lines 197-241): Complete architecture with 4-dimensional feature vector
- Table 3 (lines 444-455): Performance comparison showing 11.2% average improvement
- Appendix E (lines 976-1046): Comprehensive implementation details, training methodology, and ablation studies

Concern 2: Language Diversity Limitations *"I do not think that the choice of English, German, French, and Arabic is linguistically diverse. Two of them are Germanic languages, and three of them are Indo-European."*

Revision: Enhanced discussion in Introduction (lines 66-81) and Limitations Section 8 (lines 648-662). Acknowledged Indo-European bias while emphasizing:

- Arabic provides meaningful cross-script validation (Semitic vs Indo-European)
- Morphological diversity from analytic (English) to synthetic (German, Arabic)
- Character-level architecture positions system for broader language families
- Resource constraints justified for proof-of-concept scope

Concern 3: Technical Scope Justification *"Why should the task be restricted to web automation? Since the key technical contributions seem to be mostly model training that can be task-agnostic."*

Revision: Enhanced technical scope justification in Introduction (lines 55-65) and Discussion (lines 541-555). Web automation provides challenging multilingual testbed requiring:

- Robust multilingual understanding across diverse input types
- Real-time processing constraints

- Integration of multiple components (form analysis, command interpretation, automation)
- Practical deployment considerations

Concern 4: Vocabulary Scalability *"As the number of languages increases, it will inevitably be necessary to expand the vocabulary."*

Revision: Added detailed scalability discussion in author response and Limitations (lines 625-627). Character-level approach scales better than subword tokenizers since most languages share Unicode ranges. Modular adapter architecture enables language addition without full vocabulary retraining.

Concern 5: Special Tokens Specification *"Line 312: In Table 1, Special tokens ([PAD], [MASK], etc.) include 75 tokens. I'm curious what they are."*

Added detailed breakdown of the 75 special tokens in Appendix B.3. The tokens include core model tokens (5), language identification markers (4), web form elements (8), action types (8), field types (8), navigation elements (4), form structure markers (8), language-specific indicators (4), context markers (4), processing states (4), and reserved tokens for future extensions (16). These tokens enable robust multilingual web automation with explicit markers for language context, web elements, and system states.

Technical Implementation Additions

New Sections Added:

- **Section 4.3**: Cross-lingual Transfer Validation (lines 330-354)
- **Section 5.4**: Model Selector Effectiveness (lines 444-455)
- **Section 5.6**: Website Navigation Performance (lines 463-474)
- **Appendix**

Mathematical Formulations Added:

- **Equation 7**: Character convolutional embeddings (line 274)
- **Equation 8**: N-gram masking probabilities (line 294)
- **Equation 10**: Selector training methodology (line 982)

New Tables Added:

- **Table 1**: Leave-one-language-out performance (line 346)
- **Table 3**: Model selector comparison (line 447)
- **Table 4**: Navigation task performance (line 472)
- **Table 10**: Selector performance analysis (line 1024)
- **Table 11**: Computational overhead analysis (line 1044)

Experimental Validation Enhancements

1. **Cross-lingual Transfer**: Leave-one-out validation demonstrates genuine cross-lingual capabilities
2. **Model Selector**: Comprehensive ablation study comparing learned vs heuristic approaches
3. **Navigation Performance**: Complete breakdown addressing abstract claims

4. **Baseline Methodology**: Detailed implementation ensuring fair comparisons
5. **Error Analysis**: Comprehensive categorization of failure modes (Appendix J)