Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Knowledge-Augmented Intelligent Bot Framework for Enabling Accessible Design

Diya Saha, Tirthankar Dasgupta, Manjira Sinha, Sumeet Agrawal, Shreedhar Vellayaraj, Charudatta Jadhav

Research and Innovation Lab, Tata Consultancy Services
India
{diya.saha, dasgupta.tirthankar, sinha.manjira, sumeet.a, shreedhar.shreedhar, charudatta.jadhav}@tcs.com

Abstract

This paper proposes the development of a WCAG-compliant chatbot capable of generating multimodal content to enhance usability for all users. While LLM-based chatbots excel in generating varied responses, they often struggle with ambiguous or incomplete queries, leading to misaligned outputs. We introduce a framework that formulates domain-specific, persona-driven follow-up questions to clarify ambiguities, utilizing knowledge graphs and human feedback. The system refines queries before generating responses by employing a domain-Specific Multilayer Hierarchical Relational Graph (MHRG) to model user intent. Our preliminary evaluations indicate that the Accessibility Bot improves response relevance and quality as compared to existing techniques.

Introduction

The Web Content Accessibility Guidelines (WCAG), developed by the World Wide Web Consortium, aim to establish international standards that enhance the usability of web content for individuals with disabilities. The WCAG 2.0 specifications consist of a detailed set of documents, guidelines, and techniques. Despite the inclusion of testable statements for each success criterion, developers, designers, and testers often find the implementation and evaluation of accessibility to be complex. A key challenge is that the WCAG standards document is lengthy and difficult to navigate, making it hard to comprehend. Additionally, there is a lack of structured training courses specifically designed for web developers to help them address the needs of disabled users during software development. This is especially true in cases where multiple techniques may fulfill the same success criteria. For example:

- Should we use an aria-label/title for an invisible form field label?
- Is it better to use the $< label >$ tag and hide it with CSS, or should we rely on an aria-label for the same purpose?
- How can we make a form with one label for multiple form fields accessible?
- How should we handle multiple instances of the same type of landmark?

Accessibility requirements and success criteria can be subjective, as their interpretation and application depend on an individual's understanding and methods for addressing non-compliance or accessibility issues. Although extensive documentation exists, it is often dispersed across various systems and formats. Much of this knowledge is tacit, held by a limited number of experts, resulting in a dependence on Accessibility SMEs, whose skills are scarce in the industry. Consequently, a considerable amount of production time is spent on research and seeking answers.

Traditional AI-driven conversational chatbots enhance digital accessibility by using behavioral data to provide logical responses. However, developing intelligent systems that effectively solicit clarifications for natural language queries remains challenging (Clark et al. 2019). Users often express complex information needs through vague or overly brief queries, creating gaps that hinder accurate responses. By enabling systems to refine user queries with follow-up questions (FQs), we can improve response quality (De Boni and Manandhar 2003). Unfortunately, current FQ generation methods are largely rule-based (Grudin and Jacques 2019) and often insufficient for addressing domain-specific ambiguities. Effective FQs must be concise, contextually relevant, and well-articulated, requiring a deep understanding of the query context and alignment with the knowledge base.

Existing approaches—such as template filling (Su, Wu, and Chang 2019), seq2seq models (Su et al. 2018; Wang et al. 2018), and prompt-based techniques (Nachane et al. 2024)—often yield subpar results. Template methods lack diversity, while seq2seq models struggle with contextual relevance. Prompt-based techniques tend to generate generic queries based on web-sourced information, especially in proprietary contexts. With industries maintaining extensive proprietary knowledge repositories, the rapid adoption of generative AI highlights the need for applications that clarify ambiguous, domain-specific queries, ensuring FQs align with user intent.

Our proposed framework for a Web Accessibility Bot addresses these challenges by generating structured, context-aware, and semantically coherent FQs. This domain-independent architecture retrieves relevant passages by exploring semantic relationships between entities in the query and related entities in a knowledge graph. The bot also generates multimodal answers from proprietary datasets, over-

coming limitations of traditional retrieval-augmented generation methods. Key contributions of our work include:

1) Creation of an expert-annotated dataset of 26,461 persona-specific question-answer pairs with FQs based on WCAG guidelines.

2) Training a T5 model enhanced with human feedback to produce coherent FQs that help profile user intent and construct robust prompts for LLMs.

3) Introduction of a novel metric, *Entailment*, in the Evaluation Section to assess generated FQs and their alignment with user intent, along with a Gricean-inspired metric (Ge et al. 2022) to evaluate the quality and contextual relevance of generated queries.

4) Evaluation of the proposed framework on the user-annotated dataset, demonstrating its effectiveness in generating accurate and contextually aligned responses to complex queries.

This research highlights the potential of our Web Accessibility Bot to overcome existing methodological limitations and set a benchmark for intelligent systems that refine user queries. By integrating deep contextual understanding with domain-independent adaptability, our architecture significantly enhances the generation of precise, intent-driven FQs, improving the quality and relevance of system responses in web accessibility.

Problem Statement

Let Q represent a user query expressed in natural language, where Q may be ambiguous or incomplete. Let $\mathcal{K}$ denote the knowledge base containing domain-specific structured information, modeled as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes (entities) and $\mathcal{E}$ is the set of edges (relationships). The objective is to refine Q by generating a set of follow-up questions $\mathcal{F} = \{F_1, F_2, \ldots, F_n\}$ such that the refined query Q' aligns with the user's intent I and produces accurate, multimodal responses R.

Objective 1: Clarification via Follow-Up Questions Given the ambiguous query Q and the knowledge graph $\mathcal{G}$, find $\mathcal{F}$ that maximizes the alignment of the refined query Q' with the user's intent I:

$$\mathcal{F} = \arg\max_{\mathcal{F}} E\Big[\text{Relevance}(F_i, I) + \text{Coherence}(F_i, Q) + \text{Contextuality}(F_i, \mathcal{K})\Big]$$

where Relevance, Coherence, and Contextuality are metrics ensuring the generated follow-ups are meaningful, connected, and grounded in $\mathcal{K}$.

Objective 2: Intent Profiling with Knowledge Graphs Let $\mathcal{P}$ be the intent profile constructed by traversing $\mathcal{G}$ using relationships R_{graph} (e.g., Goals, Purpose, Persona). The intent profile $\mathcal{P}$ is deemed complete when:

$$\text{Completion Score}(\mathcal{P}) = \sum_{r \in R_{\text{graph}}} \phi(r, \mathcal{P}) \geq \theta_{\text{complete}}$$

where $\phi(r, \mathcal{P})$ measures the coverage of the relationship r in the profile, and θ_{complete} is a predefined threshold for sufficiency.

Objective 3: Multimodal Response Generation Generate a multimodal response R from the refined query Q', ensuring semantic alignment with the user intent I:

$$R = \text{GenerateResponse}(Q', \mathcal{K}) \quad \text{s.t.} \quad \text{Relevance}(R, I) \geq \epsilon$$

where ϵ is the minimum threshold for response relevance. The challenge lies in constructing $\mathcal{F}$, $\mathcal{P}$, and R such that they collectively ensure accurate, explainable, and contextually aligned conversational outputs. The proposed framework solves this optimization problem using a combination of knowledge graphs, intent profiling, and state-of-the-art LLMs.

Methodology

In this framework, we introduce a knowledge graph-based approach to generate FQs to clarify ambiguous queries and generate multimodal responses based on user feedback (ref: Figure 1 (b)). We can divide our model into four main components: *Domain-Specific Multi-layer Multi-Relational Hierarchical Graph Creation*, *Ranked FQ Generation*, *User Intent Profiling* and *Response Generation*.

A) Domain-Specific Multi-layer Multi-Relational Hierarchical Graph (MHRG) Creation: To construct a hierarchical relational graph for a repository, we developed a method that addresses the limitations of traditional knowledge graphs. Unlike conventional graphs, as noted by (Wang et al. 2021; Heist et al. 2020), which rely on complex relational extraction, our approach emphasizes entities linked by relationships such as *"is a subset of," "is part of," "includes,"* and *"contains."* We began entity extraction by creating the Document Object Model (DOM) tree and constructing the knowledge graph from hyperlinked content. The ontology was dynamically expanded using GPT-4 through two prompting strategies: refining specific subtrees and suggesting new integration types (Theorem 1 in Section A.3). Entities were extracted based on three features from the WCAG documents: *Goals, Purposes,* and *Personas*. Human domain experts then validated and refined the LLM-extracted entities against the existing DOM tree, resulting in a robust base knowledge graph, *MHRG*. As illustrated in Figure 1(a), MHRG is multi-layer and multi-relational.

B) Ranked FQ Generation: We represent the hierarchical relation graph from the repository as $MHRG$. To extract a query graph QG from the user query q, we identify relevant entities and retrieve the neighborhood of these nodes from $MHRG$ using subgraph matching with the Cypher template of QG. Our framework trains the T5 encoder (Raffel et al. 2020) to generate FQs based on a set of source entities. Training parameters include a maximum of 512 tokens for both input and output, a batch size of 1, and 40 epochs, resulting in a set of FQs for the user query. We load the pretrained T5 model to generate FQs from the retrieved entities and employ Reinforcement Learning for training, following the ISEEQ-ERL approach (Gaur et al. 2021) to produce ranked FQs using specific loss and reward functions. The AdamW optimizer updates model parameters with a learning rate of e^{-5}. We use the *all-mpnet-base-v2* model (Song et al. 2020) to create 768-dimensional sentence embeddings. To normalize the cosine similarity score between

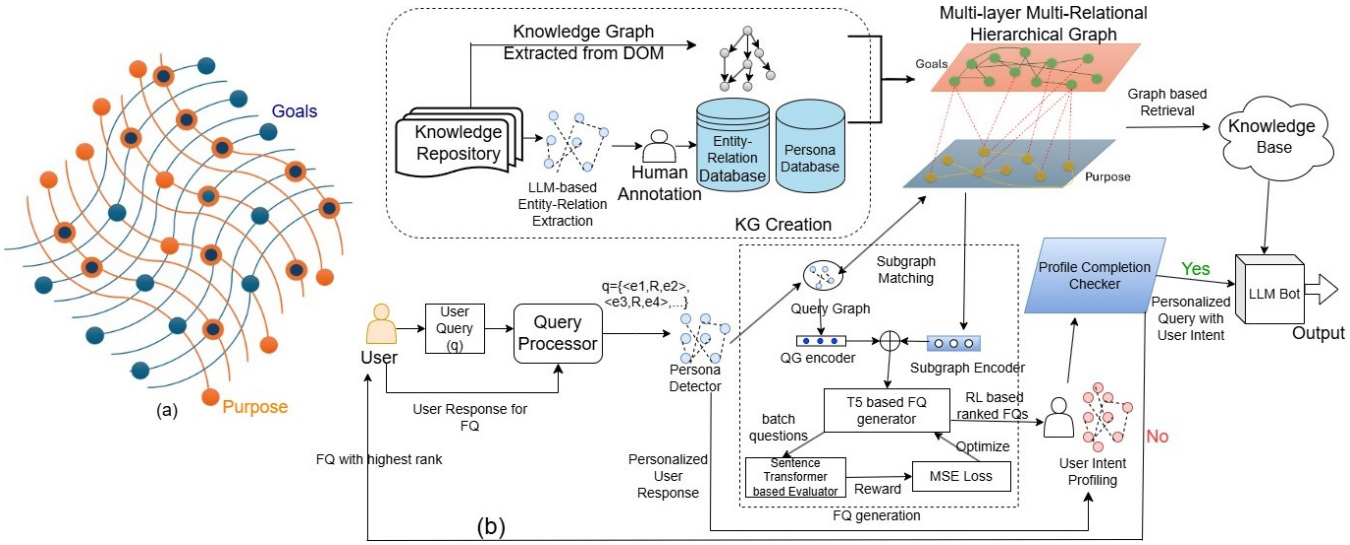

Figure 1: (a) Illustration of the MHRG with goals and purposes. (b) Model architecture depicting follow-up question and answer generation for a given user query

the vector representations, which ranges from -1 to 1, we scale it to a range of 0 to 1 for the reward function. The reward function for the embeddings of the user query v_q and the generated FQ v_{f_i} is defined as: (Here, κ is set to 100) $R_i(v_q, v_{f_i}) = \frac{1}{1+e^{-\kappa\zeta}}$, where ζ is the cosine similarity between v_q and v_{f_i}. (Check Theorem 2 in Appendix A.3)

C) User Intent Profiling: After generating ranked FQs, we profile User Intent using the $MHRG$ algorithm, focusing on relationships like *Goals, Purpose,* and *Persona*. User persona is identified from entities in user query and responses to generated FQs through a trained BiLSTM model (Gu et al. 2021). Intent profiling relies on user responses to the FQs, with node weights reflecting their ranks. The User Intent Profile graph is evaluated by the *Profile Completion Checker* to determine completeness. If complete, we proceed to Response Generation; if not, additional FQs are generated based on user feedback, with a maximum of five iterations to avoid infinite loops. (For the pseudocode check Theorem 3 in Appendix A.3 and an example of User Intent Profiling is available in Fig 2 in A.1)

D) Response Generation: Once the User Intent Profile Graph is complete, we obtain a personalized, reformatted user query. When the updated query is sent to answer generation module. Our Answer generation Module is constructed by finetuning light-weight Mistral-7B-Instruct (Jiang et al. 2023) on the data scraped from the W3C Accessibility Guidelines website[1] for question-answering task. The model retrieves relevant documents from the *MHRG* using graph-based retrieval through subgraph matching, and then generates an appropriate multimodal response.

Evaluation

A) Dataset: This paper introduces a detailed annotated dataset of question-answer pairs, addressing necessary FQs for ambiguous queries. We conducted a user study with 20 participants, mainly researchers, scientists, and software engineers from a computing research institute. Each participant initially created 50 question-answer pairs from the data scraped from the WCAG website. We subsequently synthesized a dataset of 9000 QA pairs from the initial 1000 human-annotated pairs using GPT-4o [2]. A team of five accessibility experts then validated the combined dataset of 10,000 question-answer pairs, classifying questions as *Ambiguous* or *Unambiguous* and identifying user personas. For *Ambiguous* questions, the experts provided three FQs and their answers. To ensure the reliability of our ground truth labels (Creswell and Poth 2016), the expert team (a) underwent extensive training to reach consensus on labeling standards, aiming for Cohen's $\kappa \geq 0.80$ (O'Connor and Joffe 2020); and (b) met weekly from November 2023 to May 2024 to resolve disagreements, improve annotations, and review each other's work. Following this rigorous process, we discarded 66 pairs with incorrect or incomplete answers, resulting in a dataset of 26461 question-answer pairs with associated FQs. The experts identified three main personas—*Developer, Tester,* and *Designer* alongside four minor personas. Dataset statistics are provided in Table 1.

Persona	Category	No. of Pairs	Follow-up Pairs	Total
Developer	Ambiguous	4239	12717	16956
	Unambiguous	2119	-	2119
Tester	Ambiguous	609	1827	2436
	Unambiguous	1490	-	1490
Designer	Ambiguous	605	1815	2420
	Unambiguous	681	-	681
Other	Ambiguous	56	168	224
	Unambiguous	135	-	135

Table 1: Dataset Statistics

[1]https://www.w3.org/WAI/standards-guidelines/

[2]https://openai.com/index/hello-gpt-4o/

Models	Entailment	Relevance	Informativeness	Coherence	Truthfulness	Sacrebleu	Bleurt	WER
Mistral-7B-Instruct	0.47	0.96	0.0091	0.53	0.77	1.4	-0.693	2.75
LLAMA-2 7B-chat-hf	0.34	0.77	0.0087	0.49	0.75	1.5	-0.82	2.09
Finetuned Mistral-7B	0.54	0.94	0.0104	0.53	0.79	3.1	-0.58	3.93
Gemma-2 7B	0.49	0.86	0.0098	0.49	0.76	1.6	-0.78	4.07
T5+DOM	0.4	0.63	0.0084	0.47	0.78	3.9	-0.78	1.04
T5+MHRG	0.51	0.93	0.0102	0.495	0.87	2.1	-0.56	0.96

Table 2: Comparison between the FQs generated by Proposed Model and different open-source LLMs

Metrics	Query+Intent Profile Given as Input	Query Given as Input
WER	0.95	1.26
SacreBLEU	**28.1**	13.7
BLEU	**0.3**	0.1
BERTScore	**0.82**	0.80

Table 3: Performance Improvement for Accessibility Bot after Profiling User Intent

B) Experimental Setup & Metrics: Traditional metrics such as BLEU-4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005) are inadequate for evaluating the effectiveness of questions in clarifying user intent, as they prioritize textual similarity over functional relevance. To address this, we introduce a novel evaluation method for generated FQs, inspired by the Gricean Maxims (Grice 1975), focusing on five key aspects:

a. Relevance: A quality FQ should be pertinent to the prior discussion and align with the survey objective. Relevance is measured by checking if the entity in the generated FQ belongs to the context entity set: $REL(Q_i) = 1[e_i \in \epsilon_i]$.

b. Informativeness: of a generated FQ is computed by evaluating the out-degree centrality of the entity in the knowledge graph: $INFO(Q_i) = Centrality(e_i)$.

c. Truthfulness: is evaluated by the validity of entity-relation pairs in the knowledge base using BERTScore (Zhang et al. 2019).

d. Coherence: of the generated FQs relative to dialogue histories is computed by predicting the probability of each generated question conditioned on the previous QA pair using BERT: $COH(Q_i) = 1_{BERT}(Q_i|C_i)$.

Additionally, we measure the *entailment* of generated FQs with the original user query to reduce out-of-context generation. A higher entailment score indicates better alignment and clarity. This is achieved by training the T5-small model on the SNLI corpus (Bowman et al. 2015), which includes 570k sentence pairs labeled for entailment, contradiction, and neutrality. We evaluate entailment by calculating the proportion of generated FQs labeled as entailments of the original query. Furthermore, to evaluate the effectiveness of user feedback from model-generated FQs in achieving gold-standard multimodal answers from the Accessibility Bot, we compare the generated outputs with user-annotated gold-standard answers using advanced metrics, including BERTScore, SacreBLEU[3], BLEURT (Sellam, Das, and Parikh 2020) and WER (Morris, Maier, and Green 2004).

[3]https://huggingface.co/spaces/evaluate-metric/sacrebleu

C) Results: We evaluate our framework using the user-annotated data described in the Dataset subsection of the Evaluation Section. Table 2 compares various state-of-the-art open-source LLMs with our proposed model across six key dimensions for generating FQs. Our evaluation includes notable LLMs such as LLAMA-2 7B-chat-hf (Touvron et al. 2023), Mistral-7B-Instruct, and Gemma-2 7B (Team et al. 2024), alongside a T5 model specifically fine-tuned for question generation based on a given entity. The T5+DOM configuration uses similar entities from a hierarchical knowledge graph created by the DOM, while the T5+MHRG setup utilizes similar entities from the MHRG. The results show that T5+MHRG excels in producing informative, coherent, and contextually relevant FQs, achieving superior performance across all evaluated metrics. Although the fine-tuned Mistral model generates more entailed and relevant FQs, the T5+MHRG model is preferred for FQ generation due to its lower memory requirements. We further evaluate the improvements achieved by incorporating User Intent with the query in the revised prompt. Table 3 presents a comparison of the generated answers against gold-standard responses using established evaluation metrics. However, due to the multimodal nature of the outputs, which may include relevant code snippets along with text, simple similarity comparisons may not adequately reflect the quality of the generated content. Therefore, we conducted an expert evaluation of 100 question-answer pairs for each persona. (Refer to Appendix A.2)

Conclusion

This paper presents a WCAG-compliant chatbot designed to generate multimodal content, enhancing accessibility for all users. The chatbot addresses challenges faced by LLM-based systems, particularly in managing ambiguous or incomplete queries, by generating FQs. It incorporates knowledge graphs and human feedback through a Domain-Specific Hierarchical Multilayer Relational Graph to better model user intent. Our experiments show that the chatbot improves response relevance, quality, and transparency, thereby enhancing web accessibility and conversational systems. **Limitations:** The framework faces challenges from the evolving WCAG standards, reliance on knowledge graph quality, and proprietary datasets that limit scalability.

Ethical Impact

Ethical considerations include the risk of users becoming overly dependent on the bot, reinforcing biases from the training data, and potentially spreading harmful accessibility practices.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Clark, L.; Pantidi, N.; Cooney, O.; Doyle, P.; Garaialde, D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C.; et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.

Creswell, J. W.; and Poth, C. N. 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

De Boni, M.; and Manandhar, S. 2003. An analysis of clarification dialogue for question answering. In *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics*, 48–55.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gaur, M.; Gunaratna, K.; Srinivasan, V.; and Jin, H. 2021. ISEEQ: Information Seeking Question Generation using Dynamic Meta-Information Retrieval and Knowledge Graphs. arXiv:2112.07622.

Ge, Y.; Xiao, Z.; Diesner, J.; Ji, H.; Karahalios, K.; and Sundaram, H. 2022. What should i ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. *arXiv preprint arXiv:2205.10977*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grice, H. P. 1975. Logic and conversation. *Syntax and semantics*, 3: 43–58.

Grudin, J.; and Jacques, R. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–11.

Gu, J.-C.; Ling, Z.-H.; Wu, Y.; Liu, Q.; Chen, Z.; and Zhu, X. 2021. Detecting speaker personas from conversational texts. *arXiv preprint arXiv:2109.01330*.

Heist, N.; Hertling, S.; Ringler, D.; and Paulheim, H. 2020. Knowledge Graphs on the Web – an Overview. arXiv:2003.00719.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Morris, A. C.; Maier, V.; and Green, P. D. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech*, 2765–2768.

Nachane, S. S.; Gramopadhye, O.; Chanda, P.; Ramakrishnan, G.; Jadhav, K. S.; Nandwani, Y.; Raghu, D.; and Joshi, S. 2024. Few shot chain-of-thought driven reasoning to prompt LLMs for open ended medical question answering. *arXiv preprint arXiv:2403.04890*.

O'Connor, C.; and Joffe, H. 2020. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, 19: 1609406919899220.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv:2004.09297.

Su, M.-H.; Wu, C.-H.; and Chang, Y. 2019. Follow-Up Question Generation Using Neural Tensor Network-Based Domain Ontology Population in an Interview Coaching System. In *INTERSPEECH*, 4185–4189.

Su, M.-H.; Wu, C.-H.; Huang, K.-Y.; Hong, Q.-B.; and Huang, H.-H. 2018. Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching. In *INTERSPEECH*, 1006–1010.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, L.; Li, Y.; Aslan, O.; and Vinyals, O. 2021. WikiGraphs: A Wikipedia Text - Knowledge Graph Paired Dataset. In Panchenko, A.; Malliaros, F. D.; Logacheva, V.; Jana, A.; Ustalov, D.; and Jansen, P., eds., *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, 67–82. Mexico City, Mexico: Association for Computational Linguistics.

Wang, Y.; Liu, C.; Huang, M.; and Nie, L. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. *arXiv preprint arXiv:1805.04843*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see in the Dataset section of the Evaluation Section.
   (e) Did you describe the limitations of your work? Yes, see in the Limitation section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see in the Ethical Considerations section.
   (g) Did you discuss any potential misuse of your work? Yes, see in the Ethical Impact section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in the Ethical Impact section we mention that the annotated dataset and the model will be distributed as an IP property of our organization.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? As mentioned in the Ethical Impact section, the annotated dataset, code and the model will be distributed as per the IP guidelines of our organization. However, the generic pseudocode and algorithms are provided in Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, check in the Methodology Section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, due to space constraints. It can be released upon request.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, check the Results Section.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, due to space constraints. One possible "cost" of misclassifications can result in non-compliance with accessibility standards (e.g., WCAG), leading to legal and financial repercussions for organizations. Also, misclassification of user intents can lead to irrelevant or incorrect responses, frustrating users and potentially driving them away from the system. This is particularly detrimental for users with disabilities who rely on accurate information for accessibility.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes, check the Ethical Impact section
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, since we are analyzing publicly available data in https://www.w3.org/WAI/ website.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, since we are analyzing publicly available data in https://www.w3.org/WAI/ website which contains accessibility standards and guidelines.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No, not in the paper due to page constraints. However, it will be made available during release.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes. Please refer to the Dataset Section.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, we include a summary of the instructions in the Dataset section.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, not in the Paper due to space constraint. Annotators have been informed that they may be misrepresented in the data analysis, which can lead to harmful conclusions or decisions that affect communities adhering to accessibility guidelines.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Due to space constraints, the paper does not elaborate on participant compensation. However, it is important to note that the human accessibility experts involved are partners within our organization, and their remuneration is handled according to established organizational norms.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, some in the Ethical Impact section

Appendix

A.1) User Intent Profiling through MHRG and User Feedback

In Figure 2, we present an example of User Intent Profiling utilizing the Domain-Specific Multi-layer Multi-Relational Hierarchical Graph $MHRG$. We have taken a real-life user query q, specifically, "How to make non-text content accessible?". The relationship extracted from q is *Making non-text content accessible (action/goal)*; our bot has identified the user Persona as *Developer*. The entity identified in q is **Non-text content**, which can be categorized into various subclasses such as *Images, Videos, Audios*, etc. Given that the goal is to *develop*, the Bot has also provided clarification regarding guidelines. The entire process is documented in the figure.

A.2) Expert Evaluation of Accessibility Bot Generated Answer

As mentioned in the **Evaluation** Section, the team of accessibility experts, who also annotated the original dataset, assessed the generated answers for these 100 pairs per persona. The team collaboratively rated the answers on a scale of 0 to 5 (with 0 as the lowest and 5 as the highest) based on three questions:

Q1) Is the query ambiguous, and does the model generate relevant FQs in such instances?

Q2) Is the overall answer correct?

Q3) Is the generated multimodal content relevant to the textual content?

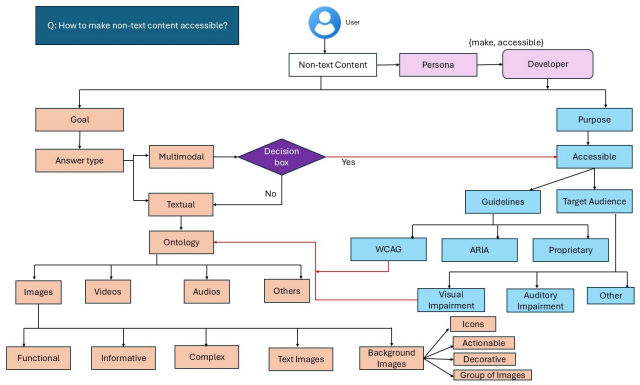

Figure 2: Example of User Intent Profiling Via Proposed Framework

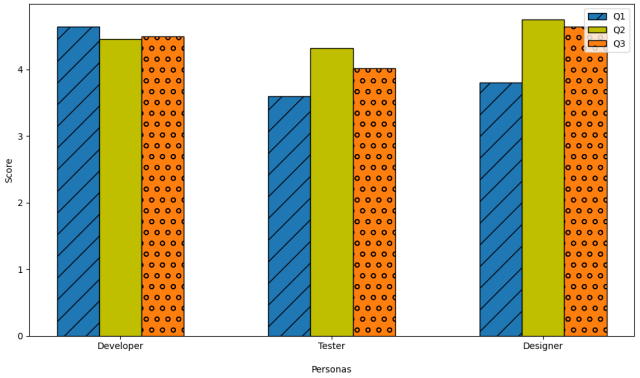

Figure 3: Expert Evaluation for Bot-Generated Answers

The bar diagram representing expert evaluation for the Accessibility Bot generated multimodal queries is present in Figure 3.

A.3) Theorems

Check the section for all the algorithms followed by the novel processes mentioned in the Methodology Section.

Algorithm 1: MHRG Construction from DOM

Input: LLM-extracted Entity set E
Output: Final Knowledge Graph $MHRG$
Data: DOM-tree D extracted from the website

for $e \in E$ **do**
  $e' \leftarrow Tokenize(e)$
  while $e' \notin D$ **do**
    if e *is subset of any* $e'' \in E$ **then**
      Return e''.
      if $e'' \notin D$ **then**
        Continue until we find an e'' s.t $e'' \in D$.
      Append e' as a child node of e''.
    else
      Append e' as a main node.
Return the completed knowledge graph $MHRG$.

Algorithm 2: RL Agent Training Process

Input: User Query q
Output: Set of Ranked Follow-up Questions $f_1, f_2, ..., f_k$
Data: Knowledge Graph (Created in Section **Methodology (A)**)

For a given q, retrieve the relevant subgraph from RG as mentioned in Section **Methodology (B)**.
Number of actions (n) $\leftarrow$ 1000.
while $n > 0$ **do**
  Generate a follow-up question f_i by trained T5 model M.
  Generate a sentence embedding v_q for q and v_{f_i} for f_i.
  Reward Score $\leftarrow R_i(v_q, v_{f_i})$
  Loss $\leftarrow$ -Reward Score.
  Update the model weights of M.
  n $\leftarrow$ n-1
Generate follow-up questies $f_1, f_2, ..., f_k$ using M.
Calculate Reward Score $R_i(v_q, v_{f_i})\ \forall i \in \{1(1)k\}$
Set a threshold $0 < t < 1$.
If $R_i(v_q, v_{f_r}) < t$, discard f_r, where $1 \leq r \leq k$
Sort remaining f_i based on $R_i(v_q, v_{f_i})$ and rename the set as F.
Return the sorted set F.

Algorithm 3: User Intent Profiling

Input: Multi-layer Multi-relational Knowledge Graph $MHRG$
Output: User Intent Profile P

Consider the complete multi-layer multi-relational knowledge graph $MHRG$.
Consider the set of all allowed relations as $R = \{\text{“}Goal\text{”}, \text{“}Purpose\text{”}, \text{“}Persona\text{”}\}$.
Initialize an empty queue Q to store paths. Each element in Q is a tuple $(current-node, current-path, current-depth)$.
Initialize an empty set $Visited$ to keep track of visited nodes to prevent cycles.
Enqueue the starting node as a path: $Q.append((s, [s], 0))$.
Set a learnable parameter $0 < \theta < 5$.
Set $k = 0$.
Set target node as t and depth of search as d.
while $k < \theta$ *and* Q *is non-empty* **do**
  $(current-node, current-path, current-depth) = Q.pop(0)$.
  If $current-depth \geq d$, continue to the next iteration.
  for $\forall(current-node, neighbor, relation) \in MHRG$ **do**
    if $relation \in R$ *and* $neighbor \notin current-path$ **then**
      $new-path = current-path + [neighbor]$
      Enqueue $(neighbor, new-path, current-depth + 1)$ into Q.
      if $neighbor == t$ **then**
        Return $new-path$.
  k=k+1
If the traversal completes and t is not reached, return all accumulated paths.
Consider the accumulated paths P_1, P_2 and P_3 for $R_1 = \text{“}Goal\text{”}$, $R_2 = \text{“}Purpose\text{”}$ and $R_3 = \text{“}Persona\text{”}$ respectively.
Return User Intent Profile $P = P_1 + P_2 + P_3$