



Multi-Document Non-Factoid Question Answering

A thesis submitted in fulfilment of the requirements for the
degree of Doctor of Philosophy

Valeria Baranova

M.S.E. in Computer and Information Science, Ural Federal University

B.S.E. in Mathematics and Computer Science, Ural Federal University

ORCID: 0000-0002-0578-3978

School of Computing Technologies
College of Science, Technology, Engineering and Maths
RMIT University

Australia

July 2024

Declaration

I certify that except where due acknowledgement has been made, this research is that of the author alone; the content of this research submission is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

In addition, I certify that this submission contains no material previously submitted for award of any qualification at any other university or institution, unless approved for a joint-award with another institution, and acknowledge that no part of this work will, in the future, be used in a submission in my name, for any other qualification in any university or other tertiary institution without the prior approval of the University, and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of any published works contained within this thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my research submission to be made available on the web, via the University's digital research repository, unless permission has been granted by the University to restrict access for a period of time.

Valeriia Baranova
24 July 2024

Acknowledgements

My PhD research journey turned out to be one of the most interesting and rewarding experiences in my life, providing the opportunity to choose not only the solutions and approaches but also the problem space itself. I have been incredibly fortunate to be surrounded by the most supportive people: my supervisory team, colleagues, and family. I am deeply grateful for their constant support, which helped me overcome various challenges, from global pandemics and geopolitical events to the joys and trials of becoming a parent. Their encouragement and trust enabled me to achieve things I never imagined possible. As my best friend's favorite quote says, "We do not have problems, we have solvable tasks," a sentiment that now extends to all areas of my life beyond research, motivating me to set high goals and learn from mistakes.

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Mark Sanderson and Prof. Falk Scholer, for their guidance, mentorship, encouragement, and immense trust in me throughout my research journey. Through our discussions and brainstorming sessions, I have grown not only as a researcher but also in my ability to trust and delegate. Their confidence in my capabilities allowed me to undertake ambitious projects and meet tight deadlines, fostering an environment where I could thrive.

Mark, your unwavering support and trusting approach have been crucial to my journey. From your first email, I knew I would be in the best hands if I could make it to Australia for my PhD, and I was not wrong. Your undeniable experience and honesty, even your willingness to review my work late at night and support my many ventures, were invaluable. Your encouraging and extremely supportive attitude, which was completely new to me, has taught me the importance of trust and delegation, enabling me to achieve goals I never thought possible.

Falk, your constructive criticism and structured guidance were essential to my success. Your ability to balance support with critical feedback helped me grow as a researcher. I am grateful that we share common interests beyond research, which made it easier to connect and build trust. I am truly privileged to have you both as my supervisors.

I am also grateful to Professor Bruce Croft, whom I was privileged to have as my third supervisor. His motivation and ideas from our first meeting shaped the structure of my research. His guidance and experience transformed my initial ideas into fundamental and pivotal research, enabling me to cover my research area comprehensively.

I would like to extend my gratitude to my collaborators: Mr. Yukun Zheng, Dr. Berkant Barla Cambazoglu, Sofya Filippova, and Dr. Luke Gallagher for their invaluable contributions to this study. I thank Mr. Yukun Zheng for conducting the eye-tracking user study, managing participant interactions, and collecting the eye fixations. Thanks to Dr. Berkant Barla Cambazoglu, who significantly contributed to the experimental design and data collection, serving as the senior study coordinator for answer utility annotations. He also provided tremendous support and companionship throughout this journey. I thank Ms. Sofya Filippova for assisting with the preparation of deteriorated data, including automatic augmentations of factual inconsistencies and manual revisions. Finally, I thank Dr. Luke Gallagher for providing insightful comments for the benchmark collection we created and contributions that greatly enriched the work.

I would also like to thank my colleagues, Dr. Oleg Zendel and Dr. Leila Tavakoli, for their friendship and for keeping me connected with the university's requirements and providing support during the longest lockdown in the world and my maternity leave.

I extend my heartfelt appreciation to my family – my parents, Larisa and Valerii, and my sister, Daria. My parents, in particular, sacrificed a lot for me during my entire educational journey. Especially during my PhD, not only did they allow me to be on the other side of the world, but they also moved across the globe themselves for almost a year to help us with home duties and babysitting so I could focus on my research. I am incredibly fortunate to have you behind my back, providing endless support, and it warms my heart to think I can make you proud.

I also want to thank my son, Leon, who was born during my PhD journey. Despite being just over two years old, your curiosity, smile, and innocence provide endless motivation and inspire me to achieve great things while reminding me to enjoy the present moments. Being your mom is the best thing that has happened to me and the greatest motivation of all.

Last but not least, I want to extend my deepest gratitude to my favorite person in the whole world, my best friend and beloved husband, Vladislav. I truly cannot find the words to express how much you have done for me and my research, particularly as my colleague and partner. You are the person with whom I generate the best ideas and bring them to life. You participated in all the brainstorming sessions almost daily, reviewed experiments, paper content, and sometimes even worked on the experiments together. You always found the right words to support me, pushed me further in the right way, and helped me with anything I needed. I can confidently say that these achievements would not have been possible without your constant support. You are my source of inspiration and my second half, and for that, I am forever grateful.

The research in this thesis was supported by the Australian Research Council (grant DP180102687), and I am grateful to my examiners for their time, constructive feedback, and valuable insights.

Contents

Declaration	ii
Acknowledgements	iii
List of Figures	ix
List of Tables	xi
Abstract	1
1 Introduction	3
1.1 Motivation	4
1.2 Objectives and Contribution of Study	6
1.3 Thesis Structure	13
2 Background	15
2.1 Brief Overview of QA Systems	16
2.2 Non-Factoid Question Answering	17
2.2.1 NFQA Related Tasks	17
2.2.2 Evolution of NFQA Systems	18
2.3 Taxonomies of Questions Types	20
2.4 NFQA Benchmarks	23
2.5 Answer Evaluation in NFQA	27
2.5.1 Answer Presentation	28
2.5.2 Eye-tracking and User Search Interaction	29
2.5.3 Answer Aspect Taxonomies	30
2.6 Chapter Summary	31
3 Exploring NFQA Evaluation Through Eye-Tracking: Human vs. Neural Network Attention	33
3.1 Introduction	34

3.2	User Study	34
3.2.1	Tasks and Users	35
3.2.2	Procedure	36
3.2.3	Experimental Setting	38
3.3	Analysis of Answer Evaluation	38
3.3.1	Overall Answer Quality	38
3.3.2	Answer Quality Aspects	39
3.3.3	Gaze Analysis	41
3.3.4	Words Annotated While Answer Rating	42
3.4	Impact Of Highlighting Words	45
3.4.1	Assessment Quality	46
3.4.2	Important Word Annotation	47
3.4.3	Differences In Gaze Metrics	47
3.4.4	Similarity of Highlights and Annotations	49
3.5	Analyzing Transformer Attention	50
3.5.1	Attention Construction	51
3.5.2	Comparison With User Annotations	52
3.5.3	One Head Is All You Need	52
3.5.4	Implications of Highlighting	53
3.6	Discussion	54
3.7	Chapter Summary	55
4	Aspects of Human-Perceived Answer Utility in NFQA	57
4.1	Introduction	58
4.2	Taxonomy of Answer Utility Aspects	59
4.3	Editorial Study	61
4.3.1	Study Procedure	61
4.3.2	Pre-study Questionnaire	62
4.3.3	Task 1: Question Generation and Labeling	63
4.3.4	Task 2: Answer Generation and Labeling	65
4.3.5	Task 3: Answer Utility and Aspect Labeling	68
4.3.6	Post-study Questionnaire	69
4.4	Answer Utility Aspect Analysis	70
4.4.1	Key Aspects of Perceived Usefulness	70
4.4.2	Factors Influencing the Perceived Usefulness of Answers	73
4.4.3	Comparison of Human and System-Generated Answers	74
4.4.4	Answer Quality Metrics Evaluation	75
4.5	Chapter Summary	77
5	A Non-Factoid Question-Answering Taxonomy	80
5.1	Introduction	81
5.2	Non-Factoid QA Taxonomy	82
5.2.1	Bootstrapping of Categories	83
5.2.2	Analysis of QA Process	83
5.2.3	Category Labelling	85
5.2.4	Taxonomy Validation	87
5.3	Question Category Prediction	90

5.3.1	NF-CATS Dataset	90
5.3.2	Question Category Classification	93
5.4	Benchmarking Categories on QA Datasets	94
5.4.1	Analysis of Category Distribution	94
5.4.2	NFQA Model Performance Across Categories	96
5.5	Chapter Summary	98
6	WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering	99
6.1	Introduction	101
6.2	Resource Description	105
6.2.1	Data Collection	105
6.2.2	Dataset Statistics	107
6.2.3	Quality Verification	109
6.3	Baseline Models	114
6.4	How To Evaluate How-To Answers	115
6.4.1	Evaluating Human Evaluation	116
6.4.2	How Good Are The Baselines, Really?	122
6.5	Chapter Summary	125
7	Conclusions and Future Work	126
7.1	Summary of Contributions	128
7.2	Future Research	130
Bibliography		133
Appendix A List of Publications		169
Appendix B Ethics Approval		171

List of Figures

2.1	Timeline of QA research milestones.	16
3.1	Procedure of the user study.	36
3.2	Example of highlighted answer display.	37
3.3	Example of the final screen used for word annotation.	37
3.4	Correct/incorrect (top/bottom) answer heatmap.	42
4.1	Aspects of answer utility.	59
4.2	Editorial study workflow: Q is a question; H and S are human and system answers, respectively; and L are label sets obtained using Q, H, and S. . .	62
4.3	Three types of Google SERP snippets available for the question “What is harry potter about?”	68
4.4	Mean values of different aspect labels.	72
4.5	Mean values of aspect labels for human and system answers.	76
5.1	Taxonomy creation procedure.	83
5.2	Interface for the question clustering study.	88
6.1	ChatGPT’s response to the question "How to clean a plastic water tank." The instructions are well-structured but contain a factual inaccuracy regarding the recommended sanitization duration.	102
6.2	WikiHow’s response to the question "How to clean a plastic water tank.", specifying a correct 20-hour sanitization period and citing reputable sources.	102
6.3	An instance of developed WIKIHowQA.	104
6.4	An example WikiHow article.	105
6.5	Human-written article summary on WikiHow.	107
6.6	Category distribution of WIKIHowQA.	108
6.7	Interface for the assessment of supporting document relevance (cropped from the right to fit the page for better visibility).	112
6.8	WIKIHowQA quality annotation.	113

6.9 Interface for the first trial of the factual inconsistency identification study (cropped from the right to fit the page for better visibility).	118
6.12 IDR by type.	120
6.10 Interface for the second trial of the factual inconsistency identification study (cropped to fit the page for better visibility).	121
6.11 IDR by categories.	122
6.13 Interface for A/B human evaluation - Top part.	123
6.14 Interface for A/B human evaluation - Bottom part.	124

List of Tables

2.1	Overview of non-factoid QA datasets.	23
3.1	Three examples of questions used in the study.	35
3.2	Mean answers quality ratings (scale from 1 to 5).	40
3.3	Overlap between explicit word annotations.	43
3.4	Overlap score between implicit annotation.	44
3.5	Similarity between explicit and implicit annotations.	46
3.6	Accuracy of rating corresponding to ground truth.	46
3.7	Gaze metrics for different answer quality levels.	48
3.8	Similarity between model and user annotations.	49
3.9	Best-head Jaccard coefficient for <i>fine-tuned</i> BERT.	53
4.1	Distribution of the labels from askers.	64
4.2	Distribution of the labels from answerers.	66
4.3	Distribution of the labels assigned by the answerers regarding the estimated usefulness of their answers.	66
4.4	Sorted list of answer sources used by answerers.	67
4.5	Distribution of the perceived usefulness labels assigned by the askers for human and system answers.	69
4.6	Correlations between the perceived usefulness and aspects of answers (non-factoid and factoid questions).	71
4.7	Correlation of attributes with perceived usefulness of answers.	73
4.8	Correlation of attributes with answer aspects.	73
4.9	Various statistics about the perceived usefulness of human and system answers.	75
4.10	Correlations of quality measures with the perceived usefulness of answers and their aspects.	77
5.1	The proposed taxonomy of NFQ categories and target answer structures.	82
5.2	Examples of evaluated questions and human/system answers from the editorial user study.	84

5.3	Perceived difficulty of questions on scale from 0 (very likely to answer) to 2 (not likely to answer); and the usefulness of corresponding human and system answers, rated from 0 (not useful) to 4 (very useful).	85
5.4	Random examples of questions grouped by workers.	91
5.5	Breakdown of NF-CATS dataset.	92
5.6	Performance of classifier models on NF-CATS test set.	94
5.7	Question category distributions in datasets.	95
5.8	Per-category performance of a state-of-the-art model: A/B human evaluation on ELI5.	97
6.1	Comparison of WikiHowQA with other QF-MDS and LFQA datasets. .	109
6.2	Percentage of novel n-grams.	110
6.3	Training hyper-parameters for BART.	115
6.4	Decoding hyper-parameters for BART.	115
6.5	Baseline models for WIKIHowQA.	115
6.6	Examples of deterioration types.	115

Abstract

This thesis focuses on multi-document answer generation for non-factoid questions (NFQs), a challenging task of generating answers to open-ended questions that can only be answered by summarizing information from multiple supporting documents. We first conducted an eye-tracking study on the user evaluation of passage-length answers to NFQs and found that the overall answer quality depends on several aspects. To better understand what makes an answer useful, we identified 12 aspects of answer quality and evaluated the limitations of commonly used NFQA evaluation measures.

To enhance understanding of the categories of NFQs that people tend to ask, what form of answers they expect to see in return, and what the key research challenges of each category are, we developed the first comprehensive NFQA taxonomy constructed with a transparent methodology and extensively evaluated via crowdsourcing. Our analysis of major NFQA datasets shows that the most challenging NFQ categories are underrepresented.

To address this, we introduce WikiHowQA, a new multi-document NFQA benchmark built on WikiHow, a website dedicated to answering “how-to” questions. We provide strong baselines and propose a novel human evaluation framework that utilizes highlighted relevant supporting passages to mitigate issues such as assessor unfamiliarity with the question topic.

We aim for this research to serve as the basis for redefining user interactions with search engines, ultimately creating the answer retrieval systems users have long desired.

Keywords: non-factoid question answering

LIST OF TABLES

Chapter 1

Introduction

As technology transforms our daily lives, from virtual assistants to advanced search algorithms, the importance of Question Answering (QA) systems has greatly increased because they address user needs directly, unlike classical search systems that merely link to relevant content [37]. These systems help us access a wealth of information, and the typical task of a QA system is to provide a useful answer to a user’s question. But how do these systems generate such answers, and how do we evaluate their utility?

When QA systems are asked about today’s weather, for instance, a user might expect an answer detailing the temperature — a fact that is quantifiable and verifiable. This exemplifies factoid question-answering, the most widely studied task in QA [199]. Factoid answers are concise, anchored to specific data points such as numbers or dates, and their accuracy is objectively measurable.

The success in factoid QA is evident through the establishment of broad and detailed taxonomies [77, 172, 176, 180], leading to the development of comprehensive datasets like SQuAD 1.1 [154], SQuAD 2.0 [156], and TriviaQA [94]. Research in factoid QA has resulted in models achieving superhuman performance on these datasets, raising the bar for accuracy and efficiency [167]. However, this progress highlights the distinct challenges of the less-explored yet significantly more complex area of non-factoid question answering (NFQA) [24].

The recent advancements in large language models (LLMs) have revolutionized the QA field, surpassing human capabilities in factoid QA tasks requiring fact-based reasoning [69]. In contrast, NFQA involves conceptual reasoning and knowledge-grounded

synthesis—areas where humans continue to outperform even state-of-the-art LLMs [69]. While LLMs are becoming a key technology in the QA field, numerous unique challenges in NFQA remain unresolved, including the evaluation and understanding of answer forms for different types of longer responses. Additionally, new issues introduced by LLMs, such as hallucinated answers, further underscore the importance of advancing research in NFQA evaluation.

This thesis aims to fill these gaps in NFQA research by exploring user interaction with NFQA systems, developing robust evaluation metrics, and creating comprehensive benchmarks.

1.1 Motivation

Non-factoid question answering covers a broad range of queries, from procedural inquiries like ‘How can you cook poached eggs without vinegar or a poacher?’ to abstract, philosophical questions such as ‘Should art produced through artificial intelligence and human collaboration be classified as human-created?’. These questions demand answers that go beyond mere facts, requiring more context and a multi-layered understanding [79]. As these systems increasingly influence critical decision-making across various sectors, from healthcare [142] to law [68], their capacity to handle intricate and diverse questions becomes crucial.

NFQA often requires systems to interpret and integrate varied viewpoints [209]. Depending on the nature of the questions, users might seek balanced opinions and arguments or detailed factual explanations. QA systems are also expected to offer sufficient context to validate the answers [84] and ensure the reputation and trustworthiness of the sources. This expectation reflects broader trends in digital interactions, where users want not only quick responses but also deeper engagement and understanding from the technology they use. Unlike factoid QA, which excels in delivering concise, data-driven responses, NFQA answers are inherently more diverse and complex, making the answer presentation, evaluation, and justification much more intricate.

The construction and evaluation of longer, passage-level answers in NFQA pose unique challenges, requiring systems to understand questions, explore context, summarize, and reason effectively. Recent advancements in LLMs have enabled the generation of highly coherent and fluent answers [32]. However, a critical challenge lies in evaluating these answers, as they may often contain inaccuracies or ‘hallucinations’ that can mislead non-expert users.

For example, evaluating process-oriented NFQA responses, such as step-by-step guides, demands scrutiny beyond fact-checking. It involves assessing the feasibility and logical se-

quencing of the steps to ensure practicality and applicability. Furthermore, for questions on debatable topics, the challenge extends to constructing balanced arguments that incorporate all relevant viewpoints to present a fair and comprehensive answer.

Addressing these challenges requires nuanced approaches tailored to the diverse demands of non-factoid questions. NFQA systems must go beyond simply generating accurate and relevant content — they need to present answers in ways that align with user expectations and adhere to the multifaceted criteria for what constitutes a useful or optimal response. A crucial task, therefore, is establishing clear standards for what defines ideal answers to different types of non-factoid questions. This involves identifying the criteria for a high-quality, non-factoid answer and understanding how users perceive the utility of these responses.

Establishing such standards, however, is complicated by the lack of a standardized taxonomy for non-factoid questions and their corresponding target answers [24, 46, 52]. Without a clear categorization of question types and answer structures, designing evaluation criteria that are precise and meaningful becomes a major challenge. This gap has slowed down the development of NFQA systems, resulting in benchmarks that have not been verified for balance or comprehensiveness. This deficiency in a clear understanding of the categories and complexities of questions makes it challenging to fully measure or evaluate the capabilities of NFQA systems, particularly in determining whether they perform equally well on simpler and more complex queries. By categorizing non-factoid questions and their expected answers, it is possible to tailor evaluation criteria to each category, leading to more precise and meaningful assessments of system performance. Establishing such a taxonomy not only addresses the diverse complexities inherent in NFQA but also serves as a foundation for the systematic advancement of the field.

Exploring how NFQA systems construct answers is crucial, particularly in multi-document contexts. While query-focused summarization (QFS) [185] — a relatively better-studied area [139, 219, 222] — addresses scenarios where a single document provides a complete answer, such instances represent only a fraction of the challenges in NFQA. Frequently, NFQA involves situations where no single source suffices, or multiple documents present divergent views on a query. This necessitates a sophisticated integration of information, pushing NFQA systems to synthesize and reason across varied sources to generate comprehensive, well-rounded answers. Our research targets this challenging new NFQA subtask, Multi-Document NFQA (MD-NFQA), which reflects the complexities of real-world information synthesis, where answers are dynamically formulated from a multitude of perspectives and data points.

MD-NFQA, as a new subtask of NFQA, lacks essential resources, such as high-quality, large-scale multi-document datasets. This scarcity has limited both the training and im-

provement of advanced models and the comprehensive evaluation of their performance, particularly regarding grounding and factual consistency with supporting documents. In this respect, MD-NFQA's progress is similar to the related task of QF multi-document summarization, which generates summaries for various query types over multiple documents and has also been constrained by the lack of sufficient large-scale, high-quality training datasets [146].

Furthermore, while advanced LLMs have shown promise in generating coherent answers, they exhibit a higher tendency for hallucinations in multi-document summarization (MDS) settings, where the synthesis of information from diverse sources is required [14]. This challenge underscores the critical need to establish knowledge-grounded benchmarks and ensure that the questions included are answerable from the provided grounded documents, as unanswerable questions can undermine the reliability of evaluations and misrepresent system performance [115].

The exploration of NFQA and its more complex subset, MD-NFQA, represents a significant area of advancement in QA systems, extending into the fields of natural language processing and information retrieval. The issues discussed in this section form the core focus of this thesis. The following subsections will outline the specific objectives and contributions of this research, which are tailored to address the identified challenges in NFQA and MD-NFQA.

1.2 Objectives and Contribution of Study

This thesis aims to address fundamental areas crucial for the advancement of NFQA, particularly within the subdomain of MD-NFQA. The main goals include a detailed analysis of the processes involved in presenting and evaluating complex long answers, the development and establishment of a standardized and streamlined taxonomy for NFQA, and the creation of a comprehensive resource for MD-NFQA. Rather than directly focusing on LLM-based QA systems, this thesis seeks to establish a fundamental understanding of the NFQA field, including user interaction with long answers, user expectations, and approaches to evaluation. These findings provide a foundation for further exploration of NFQA, including LLM-generated responses, addressing long-standing challenges such as producing useful answers and mitigating newly emerging issues like hallucinations in LLM-generated content.

To achieve these goals, this research is guided by the following four research questions:

RQ 1: How are non-factoid answers evaluated by users, what are the key criteria influencing this process, and what constitutes an optimal NFQ answer?

Given the complex nature of the NFQA task, this thesis first focuses on understanding

the process of evaluating answers to non-factoid questions. Non-factoid questions require passage-length responses that incorporate reasoning and synthesis of information. However, existing metrics used for free-form answer evaluation are insufficient for longer, abstractive texts [35, 106]. To create better metrics, we must first understand how people interact with NFQA and the evaluation criteria used by individuals. Unlike the well-explored users' interaction with snippets on search engine results pages (SERP) [12, 42, 89, 185], the interaction with NFQA is much less researched, with only one study mainly focusing on the influence of highlighting certain words in answers [152] without considering any answer quality aspects or comparing different answer presentation styles.

This research investigates the evaluation process of non-factoid answers from a user perspective. It seeks to identify the specific criteria — such as relevance, coherence, or completeness — that users consider when assessing the quality of these answers. Additionally, the question explores whether similarities exist between the evaluation patterns of users and the processing mechanisms of neural networks. The goal is to understand how the detailed elements of an answer, including the presence of certain words or the level of detail, impact its perceived quality and usefulness.

Studying user interaction with NFQA should result in defining optimal answer criteria and establishing an NFQA aspect taxonomy. While there are some aspect taxonomies for longer answers [62, 125, 170, 223], they mainly focus on community question answering (CQA) and might not be directly applicable to the NFQA field. The definition of NFQA quality aspects will enable an investigation into how current metrics align with human evaluations. It is important to understand the performance gap, particularly for NFQA, and identify any lacking answer aspects.

These optimal answer criteria are likely to be more nuanced and subjective compared to factoid QA, where answers are concise and directly verifiable. Are there specific words or facts that different users prioritize when evaluating such answers? How does the amount of detail in an answer influence its perceived quality? Are there other aspects, such as source trustworthiness, that contribute to overall answer usefulness?

Addressing these questions is crucial both for developing more effective human-evaluation frameworks and automatic evaluation metrics in NFQA. It aims to align QA systems more closely with human judgment and adapt them to meet diverse user information needs, thereby tackling the challenges in designing user-centric NFQA systems.

RQ 2: How should non-factoid questions be categorized, and are current NFQA benchmarks suitable to fairly assess QA system performance across all question types?

Given the diversity of NFQs, it is likely that optimal answers vary in structure across different categories. Consequently, different answer structures may require different evaluation

approaches. For instance, evaluating how-to instructions might involve checking the feasibility and logical sequencing of steps [171], while assessing answers to debatable questions focus on the presence of balanced and unbiased opinions [188]. Despite the importance of such categorizations, a streamlined taxonomy for NFQs and their corresponding target answer structures is absent, creating a gap in our understanding of these categories and the nature of their ideal answers [24, 46, 52]. Addressing this gap necessitates comprehensive user studies and large-scale crowdsourcing to validate preliminary findings and establish a more definitive classification.

This research aims to develop a nuanced understanding of NFQs in terms of their categorization and the nature of their answers, focusing on both specific aspects and target structures. It is also crucial to assess whether current open-source datasets, such as ELI5 [59] or MS MARCO [140] are balanced across different NFQ categories, particularly the more complex ones. This evaluation is vital because these datasets not only serve as training data but also as benchmarks for assessing QA systems. If certain categories are underrepresented, it becomes challenging to comprehensively evaluate system performance across these categories. Additionally, more balanced test sets would enable thorough validation of current models, helping to identify and address any existing gaps.

RQ3: How can we establish a comprehensive benchmark for MD-NFQA that ensures quality, balance, and factual grounding?

This research question focuses on the development of a MD-NFQA benchmark, aiming to ensure its quality, balance, and capacity for effective factual grounding. The goal is to understand how such a benchmark can be utilized to improve the training and evaluation of QA systems, particularly in MD-NFQA contexts. It involves constructing a dataset that avoids common pitfalls like data leakage due to paraphrasing, ensures a balanced representation of question types, and facilitates the grounding of answers in source documents [106]. This question is pivotal in advancing the field of NFQA by providing a robust, open-source baseline that supports the development of systems capable of handling the nuanced complexities of MD-NFQA.

RQ4: How capable are humans in annotating the quality of NFQA?

The final objective of this thesis is to evaluate the effectiveness of humans as a gold standard for assessing the quality of NFQA. Long-form QA, such as NFQA, presents unique challenges for evaluation, as standard metrics often fail to capture the nuances of generative responses [31]. While automated evaluation methods have shown promise in simpler tasks, they struggle with long-form answers generated by LLMs, particularly when detecting hallucinations or subtle inaccuracies [98]. This makes human evaluation essential for ensuring reliable assessments in such cases.

This raises important questions: Do human evaluators share the same limitations as

traditional metrics, such as failing to detect factual inconsistencies like number swaps or negations [107]? This challenge becomes even more pronounced in the context of LLM-generated passages, where hallucinations and subtle inaccuracies are common. How can evaluation frameworks be improved to address these issues and enhance human judgment?

To investigate these questions, this research utilizes a newly established baseline for MD-NFQA, introducing controlled factual inconsistencies to test the robustness of common human-evaluation frameworks. Additionally, it compares different answer presentation techniques to explore potential improvements in evaluation accuracy. The aim is to refine and enhance the reliability of human judgment in long-form QA evaluations, ensuring that human annotations remain a dependable gold standard for assessing the quality of NFQA answers.

Building on the objectives laid out in this thesis, we now discuss the contributions made to advance the field of NFQA, particularly in understanding and improving the evaluation and classification of NFQA systems, as well as enhancing the subfield of MD-NFQA.

Addressing the first part of RQ1, Chapter 3 focuses on user interaction with answers to NFQs, resulting in the following contributions:

- **Enhanced Understanding of NFQA Evaluation:** Utilizing user annotations and eye-tracking data, we identified key patterns in how answers are evaluated in NFQA, marking the first time such validation has been done using user gaze. This highlighted the ease of identifying incorrect answers compared to confirming correctness. We also demonstrated that the completeness of an answer is as important as correctness in evaluations (see Section 3.3.1).
- **Efficiency in Answer Evaluation:** Our research showed that word highlighting in answers aids users in evaluating answers more swiftly while maintaining accuracy, streamlining the process of NFQA evaluation (Section 3.4).
- **Comparison of Human and Neural Attention:** We conducted a novel analysis comparing the attention mechanisms of BERT models with human attention, utilizing both explicit annotations and implicit eye-tracking data. This study is the first to directly compare neural network attention with human gaze data. Our findings, described in Section 3.5, reveal a higher similarity between BERT model attention and human annotations than traditional methods, which are typically based on TF-IDF. This research offers new insights into the interpretability of neural networks, aligning them with human cognitive processes.
- **Novel Highlighting Model:** We proposed and made publicly available¹ a new

¹<https://github.com/Lurunchik/non-factoid-answer-highlighting>

algorithm that uses BERT model weights to highlight words, which was found to align more closely with user annotations than baseline methods. This technique can be used to enhance the presentation of NFQ answers, potentially improving both user interfaces and the frameworks used for annotations.

The research presented in the Chapter 3 resulted in the following publication:

- Bolotova, V., Blinov, V., Zheng, Y., Croft, W. B., Scholer, F., Sanderson, M. (2020). Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. Proceedings of the 29th ACM International Conference on Information Knowledge Management (CIKM '20), 85-94. <https://doi.org/10.1145/3340531.3412043>.

To fully address RQ1, we continue our study of NFQA evaluation in Chapter 4, making the following contributions to understanding useful NFQ answers:

- **First Answer Aspect Usefulness Taxonomy for NFQ:** We proposed the first taxonomy specifically designed to capture the various aspects that contribute to the usefulness of an answer in NFQA (see Section 4.2). This taxonomy, developed through an editorial study rather than crowdsourcing, provides a structured way to analyze answer effectiveness, enhancing its precision and reliability.
- **Novel Answer Aspect Dataset:** Through a detailed editorial study, we developed a labeled dataset to explore the correlation between various aspects of answers and their perceived usefulness. This dataset, now open-sourced², serves as a valuable resource for the community, allowing researchers and developers to validate and further investigate the importance of different answer aspects in NFQA.
- **Assessing Common Quality Measures:** We evaluated the effectiveness of four commonly used answer quality metrics — ROUGE, BLEU, METEOR, and BERTScore — in capturing the overall usefulness of answers. Our analysis in Section 4.4.4 provides insights into how well these metrics align with human judgments of answer quality, highlighting potential areas for improvement in automatic evaluation techniques.

The findings from Chapter 4 led to the following publication:

- Cambazoglu, B.B., Bolotova-Baranova, V., Scholer, F., Sanderson, M., Tavakoli, L., Croft, W.B. (2021). Quantifying Human-Perceived Answer Utility in Non-factoid

²Answer utility dataset is available at <https://github.com/barla/AnswerUtilityDataset>

Question Answering. In Scholer, F., Thomas, P., Elsweiler, D., Joho, H., Kando, N., Smith, C. (Eds.), CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021 (pp. 75-84). ACM. <https://doi.org/10.1145/3406522.3446028>.

In Chapter 5, we address RQ2 by investigating NFQA categories, leading to the following contributions:

- **First Streamlined Taxonomy for NFQ Categories and Answer Structures:** We introduced the first streamlined taxonomy specifically designed for categorizing non-factoid questions (NFQs) and their corresponding target answer structures (Table 5.1). This taxonomy was refined through a controlled editorial user study, which also shed light on the most challenging categories for humans to answer and offered a comparative analysis of system- and human-generated answers across different categories. The utility and accuracy of the taxonomy were rigorously tested through crowdsourcing studies. These studies included experiments on how individuals naturally group questions without a pre-defined taxonomy, providing robust validation of the new taxonomy's relevance and intuitive alignment with user perceptions.
- **NF-CATS Dataset and Classification Model:** We released a dataset of categorized NFQs alongside a well-performing classification model³. This release provides researchers and practitioners with an essential tool for quickly understanding the distribution of non-factoid question types within their datasets or for filtering questions by specific categories, facilitating effective analysis and application of NFQ data.
- **Analysis of NFQ Category Distribution in Popular Benchmarks and SOTA Model Performance:** We conducted a thorough analysis of NFQ category distribution within popular QA benchmarks and assessed the performance of state-of-the-art QA models per category. This analysis revealed that existing datasets exhibit a skewed representation across our newly developed taxonomy, particularly under-representing complex categories like DEBATE and EXPERIENCE (see Section 5.4). This gap underscores the urgent need to develop new datasets that encompass a wider range of NFQ categories and to enhance NFQA models capable of addressing these more challenging and previously underrepresented categories. This insight is crucial for guiding future research and development efforts in the NFQA field.

The research presented in this chapter resulted in the following publication:

³<https://github.com/Lurunchik/NF-CATS>

- Bolotova, V., Blinov, V., Scholer, F., Croft, W. B., Sanderson, M. (2022). A Non-Factoid Question-Answering Taxonomy. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1196-1207). ACM. doi: 10.1145/3477495.3531926

Finally, Chapter 6 addresses RQ3 and RQ4 by presenting research on establishing a new MD-NFQA benchmark and utilizing it to study and improve evaluation by human evaluators. This chapter is based on the following publication:

- Bolotova-Baranova, V., Blinov, V., Filippova, S., Scholer, F., Sanderson, M. (2023). WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5291-5314). Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.290

The following contributions have been made from this research:

- **Comprehensive MD-NFQA Benchmark:** We introduced WikiHowQA⁴, a comprehensive resource for MD-NFQA derived from WikiHow, featuring passage-level answers to instructional questions. This dataset includes parsed content from relevant HTML documents, addressing a significant gap in available resources for this type of QA and providing a crucial testbed for future research. Alongside the dataset, we conducted an analysis of human annotator performance to establish baseline benchmarks.
- **Analysis of Human Evaluators' Performance for NFQA:** In a focused study, we analyzed human evaluators' ability to assess the quality of long answers using the WikiHowQA dataset, specifically introducing factual inconsistencies to examine the accuracy of human judgment. This analysis helped to identify how reliably humans can detect errors within complex answers, providing crucial insights into potential improvements for the evaluation framework for annotators in non-factoid QA contexts, aiming to enhance the reliability and accuracy of human assessments in this challenging area.
- **Novel Evaluation Framework for Long-form QA:** We proposed and implemented a new information-augmented human evaluation framework designed to improve the reliability of NFQA annotations. This framework enhances the accuracy and utility of assessments by providing evaluators with relevant aiding passages that are highlighted based on insights gained from an eye-tracking study.

⁴The dataset and leaderboard are available at <https://lurunchik.github.io/WikiHowNFQA/>

Collaborative Efforts

Some parts of this thesis are the result of collaborative efforts with various researchers:

- **Chapter 3:** This chapter includes collaborative work with Mr. Yukun Zheng, a master's student from Tsinghua University, who conducted the eye-tracking user study, managed participant interactions, and worked with the eye-tracking software. The author of this thesis then used the raw anonymized dataset of user gaze coordinates for further analysis, model training, and comparisons.
- **Chapter 4:** This chapter was developed in collaboration with Dr. Berkant Barla Cambazoglu, who significantly contributed to the experimental design and data collection. Dr. Cambazoglu also served as the senior study coordinator for answer utility annotations. The author of this thesis participated in the experimental design and the dataset creation and then used it for subsequent analysis and metrics evaluation.
- **Chapter 6:** Sofya Filippova, a bachelor student at RMIT University, assisted with the preparation of deteriorated data, including automatic augmentations of factual inconsistencies and subsequent manual revision to filter high-quality instances. This work was designed and supervised by the thesis author, and the resulting data was used to quantify the ability of annotators to identify factual inconsistencies.

The collaborative efforts and contributions of these researchers were invaluable to the success of this research.

In general, the major tasks and contributions of this thesis, as summarized above, are detailed in the subsequent chapters. These contributions advance not only the theoretical understanding of NFQA but also enhance practical applications by providing useful datasets, tools, and improvements in the design and implementation of QA systems. The insights gained from this research facilitate future developments in both academia and industry, influencing the methods used to evaluate and enhance NFQA systems.

1.3 Thesis Structure

The rest of the thesis consists of 6 chapters and is organized to address the challenges in NFQA and MD-NFQA. Each chapter builds on the previous one to explore different aspects of the fields.

Chapter 2 — Background: We start by outlining the evolution of QA systems, focusing on the specific challenges of NFQA. This chapter introduces the existing taxonomies of question types, available NFQA benchmarks, and datasets relevant to the MD-NFQA

subtask. It also covers NFQA evaluation, including metrics and user interaction with NFQA. This chapter sets the stage for our research by identifying gaps in current studies.

Chapter 3 — Exploring NFQA User Interaction and Evaluation: This chapter examines how users interact with passage-level answers to non-factoid questions, using eye-tracking to understand where users focus and comparing this with neural network attention patterns. It provides insights into the criteria users consider important for assessing answer quality and suggests ways to improve user satisfaction and the accuracy of evaluations.

Chapter 4 — Aspects of Human-Perceived Answer Utility in Non-Factoid Question Answering: Building on the findings from the previous chapter, this chapter introduces a comprehensive list of aspects that define answer utility in NFQA. We analyze and validate a taxonomy of twelve different aspects that contribute to the value of an answer, investigating the effectiveness of commonly used answer quality measures in this context.

Chapter 5 — A Non-Factoid Question-Answering Taxonomy: This chapter develops a streamlined taxonomy for categorizing non-factoid questions and defining the expected answer forms for each category. The taxonomy is validated through extensive user studies and crowdsourcing, providing a nuanced understanding of NFQs and their corresponding answers. The chapter examines the distribution of different categories in NFQA benchmarks and evaluates current state-of-the-art NFQA systems across different categories, highlighting existing gaps.

Chapter 6 — A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering: In this chapter, we introduce the WikiHowQA dataset, a new benchmark specifically designed for MD-NFQA. This dataset addresses the unique challenges of multi-document question answering by providing human-written answers and supporting documents. We establish baseline performances and study how human evaluators assess convincing but factually incorrect model answers. We also propose a novel human evaluation framework to enhance assessment accuracy and demonstrate the remaining challenges in evaluation.

Chapter 7 — Conclusions and Future Work: The final chapter summarizes the key findings of this thesis, discusses their implications, and outlines potential directions for future research. It highlights the advancements made in understanding and improving NFQA and MD-NFQA, emphasizing the importance of continued exploration in these areas.

Chapter 2

Background

In this chapter, we provide the background to relevant previous work on Non-Factoid Question Answering (NFQA). We begin by briefly outlining the evolution of QA systems, including early domain-specific systems and the transformation to machine learning-driven models (Section 2.1). Next, we move to discussing the specific challenges of NFQA and their difference from factoid QA.

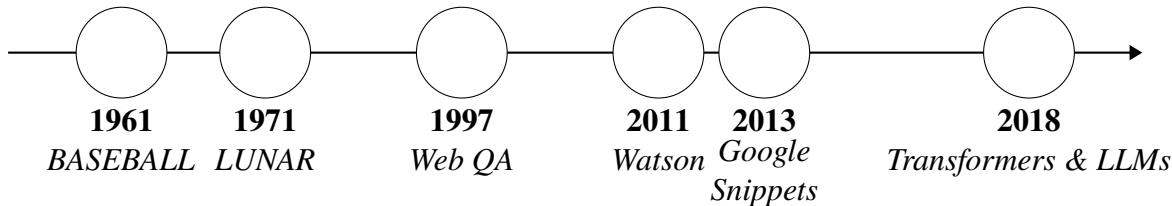
In Section 2.2, we identify and describe two primary subtasks within NFQA: answer retrieval and answer generation, discussing their respective challenges and research focus. We then provide a review of the evolution of NFQA systems. We continue by examining different taxonomies of question types, reviewing existing classifications and highlighting the need for a comprehensive taxonomy to improve system performance and evaluation (see Section 2.5.3). In section 2.4, we give an overview of publicly available benchmarks for NFQA, noting the scarcity of resources tailored for multi-document NFQA.

Finally, we address the various metrics and methodologies used for evaluating answers in NFQA, discussing the limitations of traditional IR metrics and n-gram-based metrics, and emphasizing the importance of robust human evaluation (Section 2.5). By understanding the gaps and challenges in the field, we lay the foundation for our research aimed at improving answer evaluation, deepening the understanding of non-factoid question types, and creating new tailored benchmarks.

2.1 Brief Overview of QA Systems

Question Answering (QA) systems aim to provide answers with sufficient context to validate questions posed by users in natural language [84]. These systems have been extensively researched in fields such as Information Retrieval (IR) and Natural Language Processing (NLP) [1, 22, 84, 85, 133].

Starting from early domain-specific systems like BASEBALL [72] and LUNAR [202], QA systems have evolved significantly. Kwok et al. [112] and Clarke et al. [41] first used the web in question answering, initiating IR-based QA systems that retrieve relevant documents from a large corpus and extract potential answers [104]. QA systems began to be considered an essential part of search engines as those were criticized for not fully accomplishing the task of information retrieval, leaving users to examine each document individually to find the desired information [60]. Instead, it was proposed that search engines should return a few relevant and concise sentences as answers along with their corresponding web links [133]. This marked the transformation of web search engines from keyword-based retrieval systems to natural question-answering systems [55]. In 2013, Google introduced its first featured answer snippets, which went “beyond 10 blue links” by



- 1961:** *BASEBALL system*: First domain-specific QA system developed to answer baseball-related questions, including game dates, locations, teams, and scores.
- 1971:** *LUNAR system*: Developed for analyzing Apollo lunar samples, it pioneered QA in science and allowed geologists to query data using natural language.
- 2001:** *Web QA*: Introduction of web-based QA systems, using the web as a resource for retrieving answers.
- 2011:** *Watson*: Achieved a milestone in QA by winning the television quiz show *Jeopardy!*.
- 2013:** *Google Snippets*: Google launched direct answers in search engine results, enhancing accessibility and integrating QA into everyday search experiences.
- 2018:** *Transformers & LLMs*: BERT and GPT revolutionized QA, enabling models to surpass human benchmarks.

Figure 2.1: Timeline of QA research milestones.

providing direct answers to some queries¹. Nowadays, QA is an integral and crucial part of search engines [220].

The rise of machine learning and deep learning further transformed QA systems, from models like Watson [61] to current QA systems based on transformers [187], such as various GPT [26] and BERT-based [51] models, which surpass human performance on many QA benchmarks [102, 106, 138, 197]. Figure 2.1 provides a timeline of key milestones in QA research, summarizing the field’s progression over decades.

However, researchers have shown that models achieving super-human performance on these benchmark tasks still fail on simple challenge examples and falter in real-world scenarios [102, 106]. In our work, we focus on improving the performance of QA systems in the challenging field of non-factoid questions.

2.2 Non-Factoid Question Answering

Unlike factoid QA, where questions expect short answers, usually in the form of named or numerical entities [179], non-factoid questions require longer, typically passage-level answers and can cover a wide range of topics, from information on how to fix a phone problem to getting opinions on public issues [209]. Answers to non-factoid questions are more ambiguous and vary depending on the context of the questioner and the answerer, including their experience, knowledge, and interests [24]. These answers can extend to multiple sentences or paragraphs and often have few words overlapping with the question [79]. Despite the popularity and importance of non-factoid questions, significantly less research has focused on developing QA models for these types of questions compared to factoid questions [173, 179, 213]. Since 2018, new methods addressing factoid questions have outperformed humans on the Stanford Question Answering Dataset (SQuAD)². In contrast, non-factoid questions remain an open research problem and are considered the most challenging task in QA [46].

2.2.1 NFQA Related Tasks

The task of non-factoid question answering (NFQA) can be broadly divided into two main research problems: answer retrieval and answer generation [46].

Answer retrieval involves retrieving existing answers from a pool of documents. Different research approaches this task through several methods: *answer ranking*, which orders a list of potential answers based on their relevance; *answer selection*, where the most

¹<https://moz.com/blog/101-google-answer-boxes-a-journey-into-the-knowledge-graph>

²Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset (<https://rajpurkar.github.io/SQuAD-explorer/>)

appropriate answer is chosen from a set of candidates; and *answer extraction* also referred to as *span detection* [173], which entails extracting the relevant portions from documents.

Answer generation, also referred to as long-form QA [59], addresses more complex questions that require generating detailed, multi-sentence answers from several sources that thoroughly address the question. This is a relatively recent task, as we will explore in the next section on the evolution of NFQA systems, but it is critical for developing full end-to-end NFQA systems [46].

Answer retrieval has garnered most attention in non-factoid question answering [24], with methods ranging from those based on lexical and semantic features [64, 143] to advanced transformer-based models [79, 92, 130]. However, our research focuses on the answer generation task that require synthesizing information. Specifically, we focus on MULTI-DOCUMENT NFQA (MD-NFQA), which involves using multiple relevant documents to generate a complete and coherent answer to a given NFQ grounded within the documents. This approach ensures that the generated answer is well-supported by various sources, providing a more reliable and comprehensive response.

MD-NFQA differs from the task of long-form QA (LFQA) in several ways. LFQA involves providing long answers to open-ended questions, which may include some factoid questions. For example, the ASQA LFQA dataset [177] addresses the need for long-form answers to ambiguous factoid questions, and the Natural Questions dataset [111] contains both long and short answers for factoid questions [206]. However, MD-NFQA exclusively targets non-factoid questions.

While some LFQA datasets, such as ELI5[59], include related documents, the LFQA task does not necessarily assume that an answer is scattered across multiple documents. In many cases, the answer could be contained within a single document or there might be no documents provided. In contrast, MD-NFQA specifically targets the integration of information from multiple documents to construct a single, coherent answer.

Despite the challenges, the MD-NFQA task is essential for creating comprehensive and contextually accurate answers. Simple extraction or selection from existing documents is often insufficient to address the complexity and depth required for non-factoid questions. Therefore, our research aims to advance the field of MD-NFQA.

2.2.2 Evolution of NFQA Systems

One of the earliest attempts to move beyond factoid questions focused on specific non-factoid question types [2, 150, 189, 191]. Prager et al. [150] introduced the concept of Predictive Annotation to answer "How" questions by identifying potential answers in text and annotating them to improve retrieval accuracy. Verberne et al. [190] and Higashinaka

and Isozaki [83] expanded the scope by developing approaches to answer "Why" questions using discourse and semantic relations from a pre-annotated document collection. However, they noted that additional linguistic processing was necessary to increase recall. [189, 191]. Agichtein et al. [2] focused on handling "Who," "How," "Where," and "What" questions by transforming questions into new queries that maximize the probability of retrieving relevant answers from web search engines.

Later, Soricut and Brill [174] advanced the field of non-factoid question answering by developing a QA system using an unsupervised approach to collect a large corpus of question/answer pairs from FAQ pages. This approach allowed for the training of various statistical models to enhance answer accuracy for a wide range of complex questions. Surdeanu et al. [179] further contributed by leveraging large-scale web collections, such as Yahoo! Answers, and incorporating sophisticated linguistic features, including syntactic parsing and semantic role labeling, to enhance answer ranking models.

Further research in non-factoid QA continued to enhance answer ranking models. Jansen et al. [91] proposed an answer reranking model that integrates lexical semantics with discourse information. Their approach demonstrated a 24% relative improvement over a strong baseline, still performing at only about 30% P@1, which means correctly ranking the answer for roughly one out of three questions. Fried et al. [64] improved upon this by introducing higher-order lexical semantic models that chain direct evidence to construct indirect associations between question and answer texts, though performance gains remained modest, highlighting the ongoing challenge in NFQA ranking models. Keikha et al. [100]'s experiments showed that existing passage retrieval methods that focus on topical relevance fail to perform well for non-factoid answer passage retrieval. Following that, Yang et al. [209] formally introduced the answer sentence retrieval task for non-factoid Web queries and proposed a learning-to-rank framework with semantic and context features. This framework significantly outperformed baseline methods by addressing the vocabulary mismatch problem and leveraging sentence context for better retrieval performance.

Tan et al. [181] and Cohen and Croft [43] proposed employing a bidirectional-LSTM based deep learning framework for the non-factoid answer selection task, demonstrating substantial improvements over feature-based models and previous deep learning implementations without the need for term overlap information. A combined approach using both BiLSTM and CNN for non-factoid answer selection was experimented by Rücklé and Gurevych [161], showing that integrating these models can further enhance performance. They continued by creating an end-to-end system for NFQA with an interactive visualization of neural attention weights, enabling researchers to better understand and compare attention mechanisms within their models [162]. However, the answer selection models were found to be quite limited, as they struggled to handle complex questions effectively. These models

often failed to retrieve accurate answers, resulting in suboptimal user satisfaction [95].

The development of non-factoid QA models transitioned from traditional methods to transformer-based models [59, 79, 92, 106, 130], marking a significant advance. Soleimani et al. [173] proposed a suite of models for answer extraction from multiple documents, including BERT[51], RoBERTa[127], and Longformer[15], with Longformer achieving the highest Intersection over Union score of 73.57%. As LLMs began to generate high-quality texts, answer generation became a realistic focus for research. This approach is more practical since it is not feasible to have pre-written answers for every possible question, even as parts of existing documents, making dynamic answer generation an essential step forward.

Fan et al. [59] introduced a multi-task Seq2Seq model based on the transformer architecture, pre-trained on a large dataset of question-answer pairs. They demonstrated that their model outperformed both conventional Seq2Seq and LM baselines, highlighting its ability to generate detailed and coherent long-form answers. However, unlike factoid questions, where evaluation is still challenging but relatively straightforward since the answer is usually a short text span, evaluating non-factoid questions is much more difficult [59, 152]. While these NFQA models demonstrate high performance on benchmarks, researchers have identified problems with the benchmarks themselves as well as challenges in the evaluation process and metrics [106]. To address these gaps, our research focuses on understanding the complexities of different non-factoid question types, improving the evaluation process, and creating better benchmarks.

2.3 Taxonomies of Questions Types

In 1977, Lehnert [119] stated that to answer naturally asked questions, a system needs a theory of how people ask questions and what answer types are expected. She proposed 13 conceptual categories such as Goal Orientation, Instrumental/Procedural, Quantification, Verification, etc. Graesser and Person [70] later extended the taxonomy with five categories that provide genuine inquiries. Burger et al. [28] called for new question taxonomies, highlighting limitations of past taxonomies such as a lack of scalability for the larger scope of open-domain QA, and no actual implementations of these taxonomies due to their usage requiring question processing based on a specific knowledge representation. This underscores the need for taxonomies that are both scalable and adaptable to diverse QA scenarios. Chaturvedi et al. [33] underline that the creation of a reusable taxonomy for non-factoid questions is a difficult task requiring considerable manual efforts and is expensive.

The Text Retrieval Conference (TREC 88) QA track evaluated systems that answer

factual questions. Question category classification was an important component of such systems [195] and factoid question taxonomies emerged [77, 172, 176]. These were all based on the target answer form: person, location, date, etc. However, there was no standard hierarchy of question types. Hovy et al. [87] created a QA topology of 140 answer types by manually analysing 17,000 questions, including categories proposed for NFQs (narrative answer types). The latter were marked as tentative by the authors, who suggested that more work needs to be done to evaluate and finalize those types. Suzuki et al. [180] proposed another taxonomy for factoid questions which was a hierarchy of 150 categories derived from an analysis of about 5,011 questions in Japanese.

Li and Roth [123] proposed a widely adopted two-layer taxonomy consisting of six general question categories and fifty sub-categories, and published a dataset of 6,000 labelled questions. Their taxonomy was focused primarily on factoid questions, only partially covering NFQs in a single DESC category (description and abstract concepts) represented by 1,286 questions. In our work, we expand this category to cover all non-factoid question categories, provide more detailed descriptions of categories and target answers, and conduct a thorough evaluation. Relying on these categories while creating an initial draft of the new taxonomy, we did not use the data the authors provided in our editorial studies and model training process in order to fairly compare the two taxonomies. In Chapter 5 we provide an analysis of our question category prediction model on this dataset, and map categories from one taxonomy to the other.

Gupta et al. [74] extended Li and Roth’s taxonomy through inclusion of additional sub-categories for DESC: CAUSE & EFFECT, COMPARE AND CONTRAST, and ANALYSIS. Since the NFQA taxonomy was not the focus of their paper, the authors provide only one example question for each category, they omit descriptions of new categories or expected forms of answers, do not give enough information on the methodology used to establish the categories, and do not evaluate the taxonomy to identify overlapping categories and question coverage. We drew some inspiration from their categorization during the bootstrapping stage explained in Chapter 5 and in the process found a few discrepancies such as REASON being a subset of CAUSE & EFFECT, and DESCRIBE and ANALYSIS having a blurry boundary between them.

Motivated by the Broder [25] taxonomy of user intents (later extended by Rose and Levinson [158]), Bu et al. [27] proposed a function-based QA taxonomy obtained through manual analysis of questions asked on Baidu Zhidao. It consists of six categories: fact, list, reason, solution, definition, and navigation. Unlike Broder who surveyed Alta Vista searchers along with an analysis of system logs, Bu et al. did not detail how final categories were chosen or present an evaluation of this taxonomy in terms of consistency and clarity. Mizuno et al. [135] proposed a categorization of NFQs based on types of expected answers,

and annotated 2,064 randomly sampled QA pairs from a Japanese CQA platform using that categorization. A detailed description and evaluation of the categories were not provided by the authors. Verberne et al. [189, 190, 191] studied the WHY non-factoid question category, analyzing syntactic forms of questions and types of answers that fall within that category. They released a dataset of WHY questions. While some taxonomies were proposed for WHY-type questions during that period based on different parameters, no classifications were proposed for how and definition-type question[24].

Leveraging archives of question-answering data from Yahoo!Answers, Chen et al. [36] and Guy et al. [75] categorized possible intents of CQA users. Chen et al. identified objective, subjective, and social intents, while Guy et al. classified them into informational or conversational. The broad nature of these taxonomies and their focus on social aspects rather than on the form of questions and answers complicates their direct application in NFQA systems.

More recently, Dimitrakis et al. [52], Breja and Jain [24], and Cortes et al. [46] highlighted the gap in researching unified taxonomies in their surveys. Though there is no experimental part in their research, they based their work on previous analyses and broadly established the following categories:

- Definition: questions requiring a definition;
- How: questions requiring an instruction;
- Why: questions requiring a reason;
- Opinion: questions requiring an opinion;
- Comparison: questions requiring a comparison between entities;
- Confirmation: questions checking a fact;

While these works were done in parallel with the taxonomy research presented in this thesis, the streamlined taxonomy we propose and test can almost be mapped one-to-one with these categories. INSTRUCTION corresponds to "how", REASON to "why", EVIDENCE-BASED includes "definition" and "confirmation", COMPARISON matches "comparison", and "opinion" includes EXPERIENCE and DEBATE. This parallel work underlines the importance of taxonomy as a research topic, and the resulting easy mapping demonstrates a consensus over the created categories, validating them as a sound choice. In our work, we not only develop, describe, and validate a detailed taxonomy but also create a dataset of categories, a classifier, and assess the balance of different non-factoid questions in commonly used benchmarks.

2.4 NFQA Benchmarks

In this section, we describe the publicly available datasets for NFQA. We broadly categorize them by the target task and list them in chronological order of their release. A summary of these benchmarks is provided in Table 2.1.

Dataset	Year	Lang	Context	Questions & Answers	Size	Target Task	Question Categories
nfL6[43]	2016	EN	Yahoo! Answers Webscope L6	User queries & highest-voted answers	87k	Ans. Retrieval	NFQs
PhotoshopQuiA[54]	2017	EN	Photoshop Tutorials	User queries & manual answers	40k	Ans. Generation	why NFQs
WikiPassageQA[44]	2018	EN	Wikipedia Articles	User queries & annotated passages	3.3k	Ans. Retrieval	NFQs
DuReader[81]	2018	ZH	Documents from Baidu Search & Baidu Zhidao	User queries & manual answers	200k	Ans. Generation	Mixed FQs & NFQs
MS MARCO[140]	2018	EN	Passages from Bing search	Bing queries & manual answers	1M	Ans. Generation Ans. Retrieval	Mixed FQs & NFQs
ELI5[59]	2019	EN	Relevant passages from Common Crawl	Reddit questions & highest-voted answers	272k	Ans. Generation	NFQs
ANTIQUE[79]	2020	EN	Yahoo! Answers Community	User queries & manual answers	2.4k	Ans. Retrieval	NFQs
NLQuAD[173]	2021	EN	BBC News articles (2016-2020)	News sub-headings & body paragraphs	31k	Ans. Retrieval	NFQs

Table 2.1: Overview of non-factoid QA datasets.

nfL6: The nfL6 dataset [43] was introduced in 2016 and consists of 87,361 non-factoid questions and answers extracted from Yahoo's Webscope L6³. This dataset was created using a support vector machine to ensure that only non-factoid questions were included. The answers were filtered to include only the highest-voted response for each question. The dataset is primarily used for answer retrieval tasks.

PhotoshopQuiA: The PhotoshopQuiA dataset [54] was released in 2017 and consists of 2,854 pairs of "why" questions and answers related to Adobe Photoshop usage, collected from five Community Question Answering (CQA) websites: Adobe Forums, Stack Overflow, Graphic Design, Super User, and Feedback Photoshop. The dataset focuses exclusively on why-questions, which are challenging to answer automatically as they often require information synthesized from multiple knowledge sources. The majority, 65.5% of the answers in the dataset, are official answers given by Adobe Photoshop experts. This dataset is primarily used for answer generation tasks and is valuable for developing and evaluating Why-QA systems.

³<https://webscope.sandbox.yahoo.com/catalog.php?datatype=l>

WikiPassageQA: The WikiPassageQA dataset [44] was presented in 2018 as a benchmark for non-factoid answer passage retrieval. It includes 4,165 queries sourced from Amazon Mechanical Turk⁴, based on the top 863 Wikipedia documents from the Open Wikipedia Ranking⁵. Each Wikipedia page has multiple queries, each paired with locations of relevant answer passages within the document. The dataset is designed to address the challenge of retrieving long passages that answer the non-factoid question within topically relevant documents.

DuReader: The DuReader dataset [81] was released in 2018 and is a large-scale, open-domain Chinese machine reading comprehension dataset designed to address real-world MRC challenges. It includes 200,000 questions, 420,000 answers, and 1,000,000 documents sourced from Baidu Search and Baidu Zhidao. The dataset offers a diverse array of question types, including fact-based, yes-no, and some non-factoid question categories, such as opinion-based questions. It focuses on generating comprehensive answers that often require synthesizing information from multiple documents. Its large scale and diversity make it a valuable resource for developing and evaluating Chinese or multilingual QA systems.

MS MARCO: The MS MARCO dataset [140] was introduced in 2018 as a large-scale, human-generated machine reading comprehension dataset. It consists of 1,010,916 questions sourced from Bing's search query logs, each paired with a human-generated answer. Additionally, it includes 8,841,823 passages extracted from 3,563,535 web documents retrieved by Bing, which provide the necessary context for curating the answers. The dataset supports three main tasks: predicting if a question is answerable given a set of context passages and synthesizing an answer, generating a well-formed answer based on context passages, and ranking a set of retrieved passages given a question. Due to its large size and real-world question source, this dataset has become the de facto benchmark for evaluating deep learning methods for Information Retrieval [99].

ELI5: The ELI5dataset [59] for LFQA includes 272,000 questions from the "Explain Like I'm Five" SubReddit, where questions and answers must have a score of at least two to be included. The top-voted answer for each question is considered the correct answer, and the supporting content for each question is generated by extracting sentences with high TF-IDF similarity from the top 100 web pages that match the question from the Common Crawl corpus. The dataset was one of the first large-scale benchmarks for answer generation. ELI5dataset is now a part of the KILT (Knowledge Intensive Language Tasks) benchmark [149], which provides a unified framework for training and evaluating models across a diverse set of tasks.

⁴<https://www.mturk.com/mturk/welcome>

⁵<http://law.di.unimi.it/>

ANTIQUE: The ANTIQUE dataset [79] was introduced in 2020 and contains 2,626 questions sourced from the nFL6 dataset. Since there is no complete relevance annotation in the original data, the ANTIQUE dataset added four-level relevance annotations through a multi-stage crowdsourcing process along with expert annotation. The ANTIQUE dataset consists of 34,011 QA-pair relevance annotations and is highly useful for answer retrieval tasks.

NLQuAD: The most recent NLQuAD dataset [173] contains 31,000 non-factoid questions and long answers collected from 13,000 BBC news articles. Questions and answers are extracted from the articles’ sub-headings and the corresponding body paragraphs. NLQuAD is designed to address the answer extraction task, where the answer is a long span requiring document-level language understanding.

MD-NFQA

In this subsection, we cover datasets related to the MD-NFQA task we are targeting in our research, such as LFQA and query-focused multi-document summarization (QF-MDS). We highlight the scarcity of resources available, examine their strengths and limitations, and explain how they differ from MD-NFQA tailored resources.

One of the closest datasets is the ELI5 for LFQA, as it targets non-factoid questions and includes a set of documents from which answers can be constructed. However, the authors used automated methods for curating reference content, which may not guarantee that answers are necessarily grounded in the documents or that they are not contained entirely within one document. Krishna et al. [106] demonstrated that in the ELI5, the documents are often not utilized by models. They also identified significant train/validation overlap in the ELI5, which undermines the validity of evaluation results, and highlighted challenges with both automatic answer evaluation and human annotation for this benchmark.

In our research, described in Chapter 6, we develop a new dataset aimed at addressing these problems. Our proposed MD-NFQA dataset uses texts parsed from relevant HTML pages chosen by the author of the corresponding HTML article, ensuring a higher degree of relevance and accuracy in the supporting content. We also address issues of question overlap and enhance the reliability of evaluations through a novel human-evaluation framework that leverages supporting documents.

Nakano et al. [138] used questions from the ELI5 dataset and collected new answers from human annotators who were instructed to search for related documents and use them to construct their answers. They trained the WebGPT model on that dataset to answer long-form questions by mimicking the way humans research answers to questions online: it searches and navigates the web to find relevant pieces of information and concludes an answer based on them by citing sources for factual accuracy. The authors report that the

model answers are preferred by assessors 69% of the time compared to the best human-written answers from Reddit when evaluating on ELI5.

While the WebGPT dataset is valuable, the authors have only released the model’s answers and questions, not the supporting documents. Finally, it is worth noting that the WebGPT dataset is limited to queries from the ELI5 dataset, which was constructed from the "Explain Like I’m Five" subreddit. This dataset predominantly includes questions that fall into the categories of explanations and descriptions, as will be demonstrated in Chapter 5. In contrast, our proposed resource focuses specifically on the INSTRUCTION category.

QF-MDS datasets: The QF-MDS task [185] involves creating concise summaries in response to specific queries based on multiple documents. Unlike MD-NFQA, which focuses on non-factoid questions, QF-MDS can include factoid questions or queries not in question form, such as [Entity], [Event], etc. Consequently, not all QF-MDS datasets can be utilized for MD-NFQA. Another key difference is QF-MDS summaries do not necessitate making conclusions or inferences based on the provided information. Instead, MD-NFQA requires distinct skills and approaches, higher-level reasoning and synthesis to produce complete and accurate answers. Although resources are available for Multi-Document Summarization (MDS) [6, 45, 48, 56, 66, 105, 126, 129, 198, 212] and Query-Focused Summarization (QFS) [139, 219, 222], they remain scarce for QF-MDS.

The QMDSCNN and QMDSIR datasets target QF-MDS task [147] available on request. QMDSCNN is derived from CNN/DailyMail, featuring real summaries paired with simulated queries. QMDSIR, derived from a search engine query log, includes simulated summaries with real queries. Both datasets focus on summarization, not detailed answer generation, and contain queries that may not be questions or could include factoids. These limitations make them less suitable for MD-NFQA tasks.

Another automated approach for curating large datasets for query-focused summarization tasks is AquaMuse [110]. Queries and long answers from the Natural Questions dataset [111], and a pre-processed version of the Common Crawl corpus were used to generate this dataset. Long answers from Natural Questions and the Common Crawl corpus are encoded into sentence embeddings. Then, a similarity search is performed over the corpus and long answers to find candidate documents, similar to ELI5. This dataset supports both abstractive and extractive QF-MDS tasks and shares similar limitations on its usage for MD-NFQA. Specifically, AquaMuse contains factoid questions and queries that are not necessarily in question form, which makes it less suitable for generating detailed, non-factoid answers. Additionally, the automated methods used for curating reference content may not ensure that answers are adequately grounded in multiple documents, a key requirement for MD-NFQA tasks.

Boni et al. [20] proposed HowSumm, a QF-MDS dataset automatically constructed from WikiHow content by utilizing the referenced articles as the summarization source, the corresponding elements of WikiHow articles as the target summaries, and titles used as the queries. There are two types of QF-MDS tasks, one for methods (HowSUMM-METHOD) and one for steps (HowSUMM-STEP). While this work is similar to the dataset we present, HowSumm is directed toward the QF-MDS task, does not discuss train-test overlap, and only includes the source URLs in their dataset.

2.5 Answer Evaluation in NFQA

As we covered in this chapter, there has been significant research examining the retrieval of answers to non-factoid questions, ranking of sentences, modeling of community question answering, as well as answer selection [43, 90, 100, 145, 178, 208, 209]. The most employed metrics for answer ranking and selection are well-established and extensively researched IR metrics, such as Mean Reciprocal Rank (MRR), Precision@k (P@k), Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (nDCG), and Accuracy [46].

For answer extraction tasks, the two main evaluation metrics adapted from span detection in factoid QA tasks [154] are Exact Match (EM) and the macro-averaged F1 score. Exact Match determines if the extracted answer exactly matches the target, which can be too strict a criterion for long answers. The F1 score measures the overlap between the words in the predicted and the target answers, treating sequences as a bag of words. To address the limitations of EM and F1 for long, non-factoid answers, Soleimani et al. [173] proposed evaluating models with the Intersection over Union (IoU) score, which considers partial matches. The IoU metric evaluates the quality of predicted answer spans by measuring the overlap between the predicted span and the ground truth span, thus providing a more tailored measure for longer answers.

For answer generation in NFQA, commonly used n-gram-based metrics include BLEU [144], METEOR [13], ROUGE/ROUGE-L [124], and BERTScore [217]. BLEU is a precision-based metric that calculates the number of n-grams in the candidate that appear in the reference. METEOR aligns tokens in the candidate with tokens in the reference based on exact matches, synonyms, or stem matches and combines precision and recall into an F-measure score. ROUGE-L is based on the longest common subsequence, searching for the longest co-occurring set of tokens common to both reference and candidate. BERTScore uses contextual embeddings from BERT to evaluate the similarity between reference and candidate texts.

These n-gram-based metrics, however, are often inadequate for evaluating freeform QA, especially for more abstractive answers, necessitating better metrics that go beyond n-gram

matching [35]. Since abstractive freeform QA involves complex non-factoid questions, this thesis focuses on understanding how these metrics perform specifically for non-factoid questions and their correlation with human evaluation (see Chapter 4).

Currently, human evaluation remains the primary method for evaluating non-factoid questions and text generation in general due to the limitations of existing metrics [31]. Evaluating long-form answers is particularly challenging for humans, emphasizing the need to research how people interact with longer answers and to design simpler, more effective evaluation frameworks [106]. This thesis includes research on how non-factoid answers should be presented and the development of better evaluation frameworks for NFQA.

2.5.1 Answer Presentation

In this section, we explore research on the display of answers and user reactions to them, including the utilization of eye-tracking to understand how users interact with passage-level answers to non-factoid questions.

Before answers, there were snippets: fragments of a retrieved document that matched a user’s query. The creation and value of snippets were demonstrated over twenty years ago [185]. Later, the impact of different snippet designs on user click-through was examined [42]. Results indicated the importance of users seeing all possible query terms in the snippet so that they could judge the terms’ relationships to the content of a page. Eye-tracking was subsequently employed to gain insights into how snippets might be better constructed [12].

Building on Clarke et al. [42]’s work, Iofciu et al. [89] examined two approaches to highlighting words in a snippet: query words in bold and additional words in color. This research focused on queries with ambiguous intent, highlighting words that identified intent in the snippet. Iofciu et al. [89] used both manual and automatic approaches to word choice. Lab-based experiments with manual highlighting found that users were slower and less accurate in their clicking. Automated testing, which used query log analysis and Wikipedia disambiguation pages to identify ambiguous queries and words to highlight, found that users clicked more accurately and faster when highlighting was present.

Zhang [215] investigated snippet text highlighting effects on search user behavior using both crowdsourcing and eye-tracking data. They focused on search result snippets and found that highlighted terms improved user engagement and search performance. However, their research did not explore the complexities of evaluating detailed passage-level NFQ answers. Our work extends their findings by using eye-tracking to understand user engagement with highlighted NFQ answers, providing insights into how highlighting affects answer quality evaluation.

Qu et al. [152] considered if highlighting a suite of words in an answer would allow users to identify good or bad answers more accurately. Utilizing crowd-sourced workers, the researchers found that highlighted words appeared to influence the workers' decisions, but results were inconclusive. For instance, it did not consider what aspects of an answer contribute to overall quality or whether one answer presentation style was more effective than another.

In our work, we build upon this research by designing a more detailed evaluation and utilizing eye-tracking. This approach allows us to gain deeper insights into how users interact with non-factoid answers and how different presentation styles impact their ability to judge answer quality.

2.5.2 Eye-tracking and User Search Interaction

We detail two types of studies: those concerned with how users interact with search results and those that use eye-tracking to understand the design of snippets.

Granka et al. [71] showed that a user's gaze fixated mainly on the first and second ranks of a SERP. Lorigo et al. [128] considered the impact of re-ranking results on a user's gaze. Using eye-tracking to understand the way that users scanned a SERP allowed refinements to rank learning algorithms, which improved effectiveness [93]. Eye-tracking also informed how users interact with SERPs [86, 196].

Cutrell and Guan [47] used eye-tracking to explore the impact of the length of snippets on the speed and accuracy with which users could complete tasks. They found that longer snippets helped with informational tasks, but hindered navigational [25]. Savenkov et al. [164] showed that highlighting of terms help users find the answer faster and draw their attention to results in the lower part of SERP. Lagun et al. [114] investigated the relationship between eye gaze and the browser viewport in mobile search for answer-like results from factoid questions. Some studies researched how direct answers to frequent search queries, such as weather and news, influence user behavior and how they contribute to satisfaction [37, 200].

Wu et al. [203] studied how users interact with search results when direct answers are provided in the SERP. Their study revealed that for non-factoid questions, the inclusion of direct answers in a carousel format helped users much less than for factoid questions. This research highlights the need for more effective presentation strategies to enhance user experience for non-factoid questions.

2.5.3 Answer Aspect Taxonomies

This subsection explores different taxonomies of answer aspects, such as relevance, completeness, and correctness. It discusses the role of human evaluation in assessing these aspects and the challenges involved in ensuring reliable and accurate human judgments.

Shah and Pomerantz [170] conducted a crowdsourcing study to evaluate the quality of answers in CQA. They measured the correlation between answers' quality and 13 aspects, previously proposed by Zhu et al. [223] (informative, polite, complete, readable, relevant, brief, convincing, detailed, original, objective, novel, helpful, expert). The study was somewhat inconclusive and reported no tangible correlation between answer quality and the considered aspects. Furthermore, a classifier was trained using the 13 aspects as features in order to predict the best answers. The classifier was reported to yield worse results than a naive majority class predictor.

Our work differs from the work mentioned above in several ways. First, we focus on non-factoid question answering, instead of CQA, i.e., the problem domains are not the same. Second, we conduct an editorial study instead of a crowdsourcing study. This lets us work closely with the participants and train them better for the tasks at hand, in contrast to a crowdsourcing study involving loosely coordinated and poorly trained workers. Third, the participants in the setup of Shah and Pomerantz assess answers provided to questions asked by anonymous people, while the participants in our study assess answers given to their own questions, rendering the study more realistic. Fourth, although there is a certain overlap between the aspects used in the two works, our dependent variable is the usefulness of answers, which is different from answer quality, a relatively vague target. Finally, we report high correlation between certain aspects and answer utility, unlike the work of Shah and Pomerantz, reporting inconclusive results.

Another editorial study was conducted by Arapakis et al. [8] to assess the quality of online news articles. They used 14 different aspects to model the quality of news articles (fluency, conciseness, descriptiveness, novelty, completeness, referencing, formality, richness, attractiveness, technicality, popularity, subjectivity, sentimentality, polarity). Their problem domain as well as the considered aspects differ from ours since there is no notion of questions and answers in their problem context.

While some studies aim to understand what makes an answer high quality, most of these studies focus on a limited set of aspects and are less comprehensive than ours. Lin et al. [125] investigated the optimal verbosity of answers (e.g., phrase, sentence, paragraph, document) through a user study. Users were shown to prefer paragraph-size answers, while the reliability of the answer source and the size of the search task were not found to have a significant effect on the optimal answer size. Hart and Sarma [78] conducted

a crowdsourcing study to investigate the impact of an answerer's social reputation and the verbosity of answers on the perceived answer quality in CQA. They found that novice users are likely to judge the quality of answers mainly based on some intrinsic features of the answers, such as presentation and content, instead of relying on social cues. Also, users are more interested in factors such as thoroughness and conciseness instead of answer length. Lee et al. [118] suggested that polite answers are more likely to be perceived as high quality, pointing at a politeness bias. Fichman [62] compared the answer quality of four different CQA websites and found that retrieving answers from more sources yields more complete and verifiable answers, but does not result in more accurate answers. Qu et al. [152] conducted a crowdsourcing study to observe the interaction of users with answers retrieved from a non-factoid question answering system, showing that users react to good and bad answers somewhat differently, and they can identify good answers relatively fast.

Several works model high-quality answers using low-level features extracted from questions, answers, and other sources. Fu et al. [65] trained a model using 24 textual and non-textual features to predict the quality of answers in CQA, and found that review and user features are the most powerful indicators of a high-quality answer, while the usefulness of content features vary depending on the knowledge domain. Le et al. [116] used four different groups of features (personal, community-based, textual, and contextual) to train a model to determine what constitutes the quality of answers given in the education domain. Shah [169] predicted best answers using a model trained with features extracted from the interaction history of askers and answerers. Yao et al. [211] focused on early detection of high-quality question-answer pairs in CQA. Kucuktunc et al. [109] showed that the best answers in the business domain tend to be more neutral while those in the news domain are more positive in terms of the sentiments expressed in the answers. Hasemi et al. [80] designed a neural network architecture to predict the quality of answers in non-factoid question answering systems.

Finally, a concurrent line of research exists in the context of relevance modeling in information retrieval systems. These works investigate the meaning of relevance and the aspects contributing to it. Interested readers are referred to the works of Borlund [21], Xu and Chen [207], Zhang et al. [218], and Saracevic [163].

2.6 Chapter Summary

As demonstrated in this chapter, there are substantial challenges in the task of non-factoid QA. We began by briefly outlining the evolution of QA systems in general and NFQA in particular. We discussed key milestones and methodologies that have shaped the field and examined various taxonomies of question types to improve system performance and

evaluation.

We identified two primary subtasks within NFQA: answer retrieval and answer generation. Answer retrieval involves selecting relevant answers from existing documents, while answer generation requires synthesizing new, detailed responses from multiple sources. Our review highlighted that answer generation has received less attention in research and faces significant challenges, including the development of robust models and the creation of comprehensive benchmarks.

We provided an overview of publicly available benchmarks for NFQA, noting the scarcity of resources specifically tailored for multi-document NFQA. Additionally, we addressed the various metrics and methodologies used for evaluating answers in NFQA. Traditional IR metrics and n-gram-based metrics were discussed, along with their limitations in assessing long, complex answers. We emphasized the critical role of human evaluation for the answer generation task and the challenges associated with reliably assessing non-factoid answers, highlighting the need for better evaluation frameworks—both automatic metrics and improved human evaluation methods.

In summary, this chapter aims to provide a comprehensive understanding of the complexities and challenges of NFQA. By reviewing existing research, methodologies, benchmarks, and evaluation metrics, we identified key gaps in the field which our research addresses.

Chapter 3

Exploring NFQA Evaluation Through Eye-Tracking: Human vs. Neural Network Attention

Search Engine Result Pages (SERPs), such as those on Google or Bing, increasingly feature direct answers to user queries. Similarly, Large Language Model (LLM)-based services like ChatGPT are now widely used as QA systems, providing conversational responses to both factoid and non-factoid queries. These systems exemplify the growing shift toward delivering information directly, bypassing traditional link-based search results. However, traditional metrics often fail to capture the usefulness and factual correctness of answers to non-factoid questions, which require lengthy, passage-level responses [79, 177]. This issue is further compounded by the tendency of LLMs to generate hallucinated content, making the evaluation of such answers even more critical.

This chapter explores how people evaluate such answers, a foundational step toward developing more effective metrics. Traditionally used to study SERP layouts, eye-tracking technology offers a precise method to monitor users' gaze and engagement with content[29]. This study pioneers in the examination of human gaze attention during NFQA evaluations and in investigating whether neural networks detect similar patterns. Insights from this study not only help identify key criteria users consider in passage-level answers but also suggest ways improve user satisfaction and the accuracy of assessments.

3.1 Introduction

In this chapter, we explore the evaluation of non-factoid question answering (NFQA) using a novel approach in a controlled user study. For the first time in NFQA research, we employ eye-tracking technology to observe how users interact with complex, passage-level answers. This method allows us to capture not only explicit user feedback but also implicit gaze patterns, providing a comprehensive view of how information is processed and assessed in real-time.

Building on the work by Qu et al.—one of the few studies to explore NFQA evaluation, particularly the impact of automatic word highlighting—our research aims to address gaps in understanding the overall quality of non-factoid answers and the effectiveness of different presentation styles. We extend this research by employing fine-grained evaluation criteria through eye-tracking, allowing us to compare how different metrics and highlighting affect user interaction with NFQA. We investigate whether users focus on the same words or facts when evaluating an answer and whether the level of detail in an answer influences its perceived quality.

Our research investigates three research questions aiming to deepen our understanding of user engagement with NFQA:

1. How do people interact with detailed, passage-level answers and understand whether an answer for a non-factoid question is correct? What features of an answer determine overall answer quality, and is the evaluation process similar across users and answers of different quality?
2. Does automated word highlighting help users evaluate answer quality more efficiently and accurately?
3. By analyzing user gaze data alongside the attention mechanisms of a fine-tuned BERT model [51], we investigate the correlation between the areas where people focus and what the model considers important. Can insights this model’s attention mechanism be utilized to improve the highlighting of key words in answers?

3.2 User Study

To explore how users evaluate non-factoid answers, we conducted a user study in a controlled lab environment using an eye-tracker. This study involved collecting user behavior data as they interacted with a non-factoid question-answering system under two conditions: with and without automatic word highlighting in answers. Highlighting was hypothesized to

assist users in evaluating answer quality more efficiently. The study was conducted reviewed and approved by Human Ethics Cometee RMIT University¹. This study was conducted in collaboration with Mr. Yukun Zheng, a master’s student from Thigunya University, who was responsible for the eye-tracking data preparation and the study description is based on the relevant research [18]. This section details the methodology employed, including the tasks, profiles of participants, study procedure, and experimental settings.

3.2.1 Tasks and Users

To conduct this study, we selected 40 questions from the NFL6 dataset, sourced from Yahoo!Answers — a community-driven question-answering platform. Examples of these questions are displayed in Table 3.1. Each question was paired with either its “best” answer, as determined by the original (Yahoo!) asker of the question, or with an incorrect answer selected through BM25 scoring [175] and further re-ranked using a BLSTM model [43]. The distribution of answer lengths between correct and incorrect responses was comparable, with correct answers averaging 344.0 words ($SD=152.3$) and incorrect ones averaging 419.7 words ($SD=197.4$). We employed a previously established method to highlight significant words in the answers [152], selecting the five words with the highest TF-IDF scores and all capitalized words that did not start a sentence. The complete dataset used in this study is publicly accessible².

Table 3.1: Three examples of questions used in the study.

TaskID	Question
14579170	How to get a speeding ticket dismissed?
17172970	Why is my MRSA Staph infection continuing to relapse?
17636401	How to check for dead pixels when buying psp?

We initially recruited 32 participants, from which we selected 20 based on their visual acuity confirmed via Tobii Pro X2-60 eye-tracker, ensuring reliable eye movement data. The final group consisted of 15 females, 4 males, and 1 who preferred not to disclose their gender, all of whom were either students or university staff with proficient English skills. The age distribution included 7 participants aged 18-24, 5 aged 25-30, 2 aged 31-35, 4 aged 36-40, and 2 over 40. Each participant engaged in the study for approximately one and a half hours on average and received a \$50 USD gift voucher as compensation upon completing 40 tasks.

¹The IRB approval number is 22427.

²<https://github.com/Lurunchik/non-factoid-answer-highlighting>

3.2.2 Procedure

After an initial calibration with the eye-tracker, each participant was guided through two practice tasks to familiarize themselves with the process. They then performed forty QA tasks, each involving interactions across five different screens. The complete process is depicted in Figure 3.1.

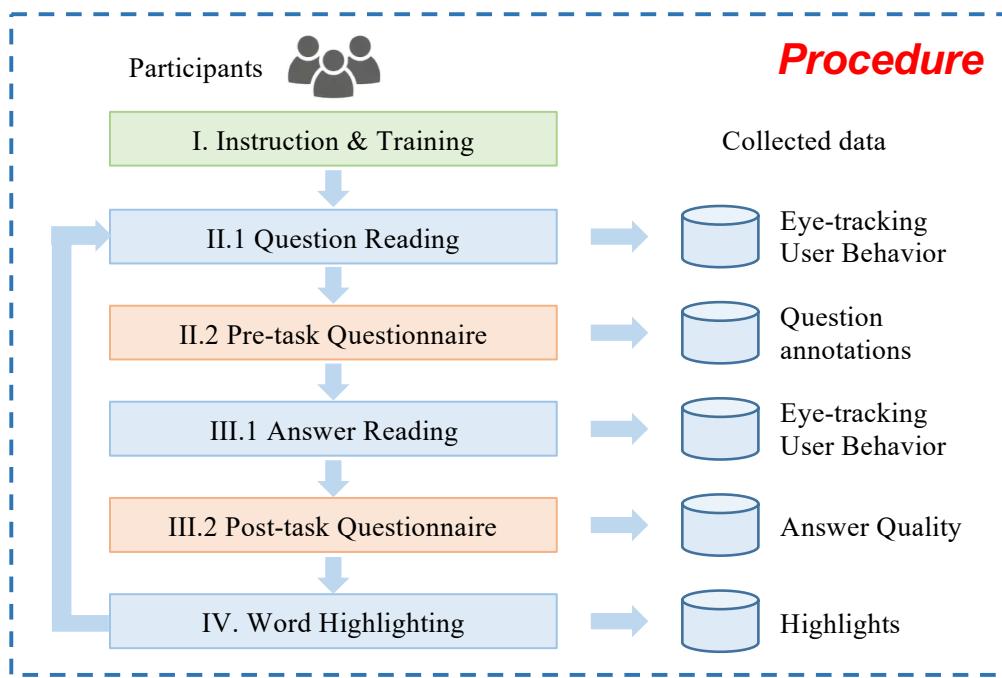


Figure 3.1: Procedure of the user study.

1. **Question Reading:** A participant reads a single question displayed on a white screen, which contains no other elements to avoid distractions. They proceed to the next screen by pressing the space bar, a method chosen to minimize the impact on eye-tracking data compared to using a clickable button.
2. **Pre-task Questionnaire:** The participant assesses their *interest*, *familiarity*, and *perceived difficulty* of the question on a scale of 1-5 (low-high), with the option to revisit the question up to two times.
3. **Answer Reading:** The participant views the entire answer, displayed with or without highlights depending on their group assignment. No scrolling is required. To proceed, the participant presses the space bar. An example of a highlighted answer is illustrated in Figure 3.2.

[Upgrade](#) to [dsl](#) and tell them what you want to do or [upgrade](#) to cable and get yourself a router. With a dial-up [connection](#), a router is a waste of time. It'll work but if you [upgrade](#) to [dsl](#), chances are it won't cost you any more (or much more) than dial-up and they'll give you all the equipment you need. [Plus](#) you'll have a much better [connection](#) with [dsl](#) over dial-up.

Figure 3.2: Example of highlighted answer display.

4. **Post-task Questionnaire:** The participant rates the *overall quality* of the given answer on a scale of 0-2, as well as its *correctness*, *completeness*, and *conciseness* on a scale of 1-5 (low-high). The quality scale is defined as: (0) an INCORRECT answer; (1) a CORRECT but low-quality (*LQ*) answer; and (2) a CORRECT and high-quality (*HQ*) answer. They may also provide additional feedback on the answer.
5. **Word Annotation:** Finally, the question and the answer are displayed again, and the participant marks positive and negative words using a mouse. They identify words that convince them the answer is correct as positive, and those suggesting the answer is incorrect or of low quality as negative. The participant can select complete words, phrases, or sentences, with guidance to focus on fine-grained marking. An example of the final screen used for word annotation is displayed in Figure 3.3.

Question: I have a dsl connection another hous member is using it for a wireless connection how can I stop access 2 min?

Passage:

[Upgrade](#) to [dsl](#) and tell them what you want to do or [upgrade](#) to cable and get yourself a router. With a dial-up [connection](#), a router is a waste of time. It'll work but if you [upgrade](#) to [dsl](#), chances are it won't cost you any more (or much more) than dial-up and they'll give you all the equipment you need. [Plus](#) you'll have a much better [connection](#) with [dsl](#) over dial-up.

Positive mode

Upgrade to [dsl](#) and tell them what you want to do or upgrade to cable and get yourself a router. With a [dial-up connection](#), a [router](#) is a waste of time. It'll work but if you [upgrade](#) to [dsl](#), chances are it won't cost you any more (or much more) than [dial-up](#) and they'll give you all the equipment you need. Plus you'll have a much [better connection](#) with [dsl](#) over dial-up.

ID	Mode	Content	Action
1	Pos	dsl	<input type="button" value="Delete"/>
2	Pos	dial-up connection	<input type="button" value="Delete"/>
3	Pos	dial-up	<input type="button" value="Delete"/>
4	Pos	upgrade to dsl	<input type="button" value="Delete"/>
5	Pos	router	<input type="button" value="Delete"/>
6	Pos	with dsl over dial-up	<input type="button" value="Delete"/>
7	Pos	better connection	<input type="button" value="Delete"/>

Finish

Figure 3.3: Example of the final screen used for word annotation.

3.2.3 Experimental Setting

Participants were randomly assigned to one of two groups. The first group evaluated answers without automatic highlighting, while the second group viewed answers with highlighted words, as previously described. For each question, its correct answer was presented to five out of ten participants, while its incorrect answer was shown to the remaining five. Thus, each answer was evaluated by ten participants in total.

3.3 Analysis of Answer Evaluation

In this section, we analyze the data collected from our user study, which includes users' prior knowledge about the questions, their ratings of answer quality, explicit markings of key words, and their eye-tracking data. We will look at how users rate the quality of answers to non-factoid questions, compare gaze metrics, and examine the differences between what users explicitly highlight and where they actually look. This analysis helps us understand how users evaluate and interact with answers.

3.3.1 Overall Answer Quality

We first consider *overall answer quality* ratings (0-2) made by users. The average agreement for ratings between all pairs of users was moderate, with a Cohen's Kappa [10] value of 0.59.

To obtain a single quality label for each answer for subsequent analysis, we took the majority vote across the ten users. In total, 43 answers were rated as INCORRECT and 37 as CORRECT; of the latter group, 22 were rated as low quality (*LQ*) and 15 as high quality (*HQ*). Recall that we constructed our data based on the initial nfL6 collection to include forty correct and forty incorrect answers; manual inspection showed the three discrepancies between our users' ratings and this dataset were due to labeling errors in nfL6 collection. These were answers rated as incorrect but marked as correct or vice-versa. *All subsequent analysis was based on this corrected grouping.*

Table 3.6 shows the accuracy of user ratings, defined as the ratio of answers labelled in accordance with the majority vote to the total number of rated answers. The accuracy is higher for incorrect answers and the most common mistake was when users rated a correct answer as an incorrect one (28 false positives versus 91 false negatives). For each answer that had at least one erroneous rating, we studied the difference between users who made a mistake and those who did not. We found the mean value of user perceived difficulty of questions was higher for users who made a mistake compared to those who didn't (2.8 vs

2.5). Mean interest and familiarity values were lower for users who made mistakes (2.7 and 2.2, respectively) vs (3 and 2.4, respectively). Note, however these differences were not statistically significant. When considering user errors, there are two types: *false positives* and *false negatives*. We will study causes of these mistakes separately.

For the answers where *false negatives* were made, there was a statistically significant increase in perceived difficulty of a question ($p = 0.04$) between users who mistakenly rated correct answers as incorrect and users who rated answers correctly. Therefore, *when a question is difficult for a user, they tend to rate a correct answer as incorrect probably because of a lack of information both in the answer and in the user's initial knowledge*.

Regarding *false positives*, there is a statistically significant decrease in the user's interest in the question between those who mistakenly rated incorrect answers as correct and those who correctly rated them ($p = 0.03$). This indicates that *users could be less attentive while rated the question they are not interested in*. Below is an example of a question-and-answer pair that received highly varied quality ratings from different users:

Q: I have a dsl connection another house member is using it for a wireless connection how can I stop access 2 him?

A: Upgrade to dsl and tell them what you want to do or upgrade to cable and get yourself a router. With a dial-up connection, a router is a waste of time. It'll work but if you upgrade to dsl, chances are it won't cost you any more (or much more) than dial-up and they'll give you all the equipment you need. Plus you'll have a much better connection with dsl over dial-up.

Here, three out of ten users rated the incorrect answer as *HQ*. All three users had an interest score of only one out of five. We hypothesise that they were misled by the high keyword overlap between the question and the answer. Notably, two of the users only saw the plain text without highlighting.

3.3.2 Answer Quality Aspects

We also obtained ratings for three quality aspects: *correctness*, *completeness*, and *conciseness*. Table 3.2 provides detailed information on the mean values and weighted Cohen's Kappa agreement between users for each aspect. The ratings suggest that even high quality answers, which were chosen as the best on the Yahoo!Answers website, can be greatly improved. There is fair agreement on the rating of correctness for incorrect answers (the mean value is 1.57 with 0.25 Cohen's Kappa). However, there is almost no agreement for the rest.

Regarding the contribution of each answer quality aspect, both *correctness* and *com-*

Table 3.2: Mean answers quality ratings (scale from 1 to 5).

	correctness		completeness		conciseness	
	mean	kappa	mean	kappa	mean	kappa
INCORRECT	1.57	0.25	1.53	0.11	1.85	0.16
CORRECT Low Quality	3.20	0.01	2.82	-0.05	2.95	0.09
CORRECT High Quality	4.25	0.03	3.93	0.10	3.54	0.14

pletteness have high correlation (Pearson, $p < 0.001$) with the overall quality rating, followed by moderate correlation for *conciseness*. The values for the three aspects were 0.85, 0.79, 0.63 measured across all answers. Separately, within only incorrect answers, the strongest correlation is correctness (0.72, 0.56 and 0.4). Within correct answers, correctness and completeness have comparable correlation with overall quality (0.79 and 0.77, 0.54).

There are few cases when correctness was assigned a high rating while completeness a low one, and we also found a high correlation between correctness and completeness (0.817) for all answers and 0.818 for only correct ones, $p < 0.001$. In comparison, correlation of correctness with conciseness was only moderate overall (0.63), and low within correct answers (0.47), $p < 0.001$. *This analysis indicates an important relationship: the answer to a non-factoid question is rated as correct only when it is also complete. For instance, the answer contains all parts of an explanation, or different opinions, examples, and so on. On the other hand, both completeness and conciseness have little meaning when an answer is incorrect and they were rated rather randomly by users.*

To understand what could influence user rating, we investigated the relationship between aspect score values and user perceptions of question difficulty, familiarity, and interest. Cases with low and high aspect score variance were analyzed separately, as sometimes the rating of an answer does not require extra knowledge, e.g. when an answer is obviously incorrect or correct. Below there is a correct Q&A pair with low score variance of answer quality (the variance is 0 overall and 0.1 for each aspects) but high variance (>2) of user familiarity with the question.

Q: Why is ice less dense than water?

A: The molecules of water are closer together and constantly moving, whereas the molecules of ice are in a crystal lattice, meaning they're in a rigid formation. When water freezes, the molecules spread out a little more to form the crystal lattice. Since density is mass over volume, and ice has takes up more volume than water, the density of ice is lesser than that of water. Which makes ice float on water.

Ideally, we want automatic QA-models to generate answers that contain enough information to be understood and assessed correctly regardless of a user's initial knowledge

about the question. For answers with higher quality score variance, we only found low positive correlation (0.24 Pearson, $p < 0.05$) between correctness ratings and interest in the question. We found a low negative correlation (-0.21 Pearson, $p < 0.05$) between difficulty and conciseness. Thus, we did not observe any definitive strong influence there.

3.3.3 Gaze Analysis

During reading, people make a series of rapid eye-movements called *saccades*, while for some periods of time the eyes are relatively still, called *fixations*. It is during the fixations that a reader acquires information [201]. We therefore focus on fixations in our analysis of our users. The text reading process also involves fixations that go against the normal reading order of left to right and top to bottom for English text. In such cases, people return back to already seen parts of a text. Such fixations are called *regressions*.

From the raw eye-tracking data, we can obtain the duration of fixations and their positions on the screen. We analyse fixations that last for more than 60 milliseconds in all subsequent analysis, following common practice [221].

Eye movements can be influenced by many factors including reading ability [11], a person's prior knowledge about a question [96], and demographics [157]. However, it has been found that people focus their attention on words that are relevant to their question [164]. We studied word-level fixations, and for each word, we identified all fixations whose coordinates fell into a word-sized bounding box. The following gaze measures were used to study how our users interacted with answers of different quality for a non-factoid question:

- *Total view seconds*, the total time spent looking at a screen with an answer.
- *Mean fixation duration*, the average length of fixations, per user and per answer.
- *Mean regression duration*, the average length of fixations which were regressions to a word, per user and per answer.
- *Mean word fixations*, the average number of fixations on a word, per user and per answer.
- *Mean word regressions*, the average number of word regressions in the answer, per user and per answer.

All results for these gaze metrics are shown in Table 3.7. The values for total view seconds, mean fixation, and regression duration are all lower for incorrect answers, which demonstrates that *users spend more time and effort to understand that an answer is fully*

correct than that it is missing some information or is incorrect. Word fixations and regressions, which have been shown in related work to be indicators of relevance [76], are lower for wrong answers. We tested the statistical significance of differences between metric values for correct and incorrect answers using ‘. Significant cases were marked in Table 3.7. Results were statistically significantly different for total view seconds, regression duration, and word regression counts. This finding is consistent with results from Qu et al. [152] that *people interact with good and bad answers differently, and rate incorrect answers with less effort.*

Typically, incorrect answers also have people gaze at more areas of interest with lower average fixation duration compared to correct answers. An illustrative example of gaze heatmaps for a house who was not shown highlighted words is shown in Figure 3.4.

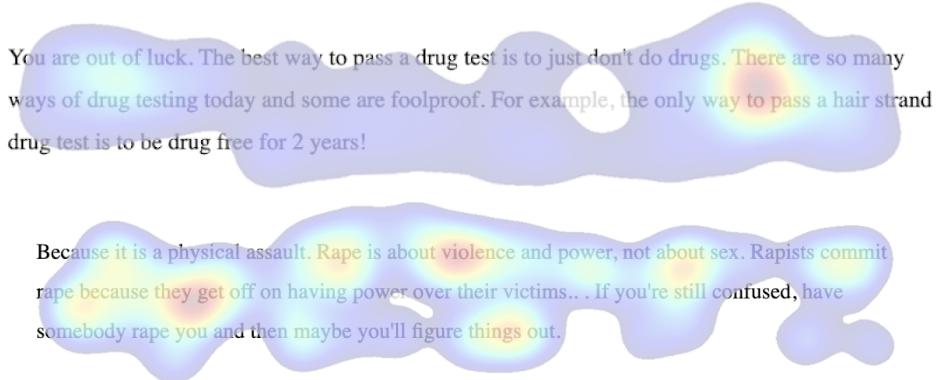


Figure 3.4: Correct/incorrect (top/bottom) answer heatmap.

3.3.4 Words Annotated While Answer Rating

For factoid questions, which typically require short answers, the important words for determining correctness might include most or all of the text. In contrast, it’s less clear which parts of a longer, passage-level answer for non-factoid questions are crucial for evaluating correctness. In our study, participants were asked to mark both positive and negative words and sections in the answers that influenced their quality assessments. Simultaneously, we recorded their gaze with an eye-tracker. This subsection examines the consistency among participants in the words they marked and those they focused on during their evaluations. We also compare the explicit annotations of participants with their gaze patterns to see how closely these two measures align.

Explicit word annotation.

Participants in our study marked words as either positive or negative based on how these words influenced their judgments about the accuracy of an answer. Notably, only a small

fraction of words (0.02%) received annotations that contradicted the final answer rating. Words in correct answers were typically marked positively, whereas those in incorrect answers were marked negatively. Most contradictions occurred when positive words were marked in answers that were incorrect. This pattern indicates that incorrect answers labeled as low quality (LQ) typically lacked necessary information rather than containing incorrect facts.

Futhermore, the relative length of annotations words in correct answers was statistically significantly higher than the length in incorrect ones. This suggests that *people found it challenging to identify exactly what was wrong in incorrect answers to non-factoid questions*. There was a statistically significant positive correlation between answer quality aspect rating and the relative lengths of annotations: 0.33, 0.29 and 0.46 for correctness, completeness and conciseness respectively (Pearson, $p < 0.05$).

To calculate the agreement between annotated spans of text in passages, we followed the approach of Qu et al. [152] and used the overlap coefficient [193]:

$$\text{Overlap}(H_1, H_2) = \frac{|H_1 \cap H_2|}{\min(|H_1|, |H_2|)} \quad (3.1)$$

where H_1 and H_2 are unique words from the annotations of two users. This metric allows us to compare annotations of different lengths as some users annotated distinct words and others whole sentences. Following Qu et al. [152], all stopwords were excluded. We calculated overlap only between positive words for correct answers, and negative words for incorrect answers because of the almost complete lack of opposite annotations.

Table 3.3 shows the mean overlap scores between user pairs, broken down by different user setting groups (columns) and different answer quality ratings (rows). Overall, the agreement for correct answers was higher, which was consistent with Qu et al. [152] and means that *users had good agreement between each other on the words that were important for the identification of a correct answer*.

Table 3.3: Overlap between explicit word annotations.

Answer quality	User settings			
	all	without highlighting	with highlighting	between 2 settings
ALL	0.66	0.68 [†]	0.70 [†]	0.64
INCORRECT	0.65*	0.66 [†] *	0.69 [†]	0.62*
CORRECT	0.68*	0.71*	0.70	0.66*
LQ & HQ	0.67 & 0.7	0.69 & 0.73	0.71 & 0.7	0.64 & 0.68

* significant difference between correct/incorrect answer groups;

† between users shown/not shown automatic highlighting
(Student's t-test, $p < 0.05$).

Implicit word annotation.

Eye-tracking data allows us to calculate agreement between words that users annotated as important implicitly, namely those that received more attention based on gaze patterns. *Implicit annotation* vectors for each user were constructed as follows. First, as a target number of annotated words for each user, we used the number k of unique words in the same user's explicit annotation. To ensure comparability despite the variation in annotation lengths across different users, we applied the same overlap measure (as defined in Formula 6.2.3) that was previously used to calculate explicit agreement. We hypothesized that words crucial to the answer would attract more gaze from users than less relevant ones. We measured gaze intensity by the amount of time that a user looked at it (the total fixation duration on each word), and the number of times users returned to it (regressions). Although individual variations might cause some users to focus longer on certain words, such as those that are rare or unfamiliar, we conjecture that words that are of interest in relation to the specific question being answered would receive more consistent attention across users.

The top k words for implicit annotation were selected by initially sorting them by decreasing regression count, and then by decreasing fixation duration to break ties. In Table 3.4, we present the overlap scores for these implicit annotations. The pattern of agreement between annotations for correct and incorrect answers similar to that of explicit annotations, with the overlap for correct answers being statistically significantly higher.

Table 3.4: Overlap score between implicit annotation.

Answer quality	User settings			
	all	without highlighting	with highlighting	between 2 settings
ALL	0.59	0.67 [†]	0.56 [†]	0.57
INCORRECT	0.58*	0.65 ^{†*}	0.56 [†]	0.55*
CORRECT	0.61*	0.70 ^{†*}	0.56 [†]	0.60*
LQ & HQ	0.62 & 0.59	0.72 [†] & 0.68 [†]	0.57 [†] & 0.54 [†]	0.6 & 0.59

* significant difference between correct/incorrect answer groups;

† between users shown/not shown automatic highlighting
(Student's t-test, $p < 0.05$).

Comparison of explicit annotation and user gaze.

To examine the overlap between words explicitly annotated by users as important for determining answer correctness and those implicitly deemed important (those fixated on and regressed to), we first removed all stopwords, as they are not relevant for this comparison.

We then assessed the overlap between explicit user annotations and implicit annotations. Given that the lists were of equal length, we employed the Jaccard coefficient as a measure of similarity. The first part of Table 3.5 displays the average Jaccard scores across all documents, calculated between the explicit and implicit word lists for each user. While the overall agreement is not particularly high, there is nonetheless a noticeable intersection of words.

As explained previously, there are many factors that can influence a user's gaze. To account for individual noise, such as long fixations on words that happen to be unknown for a particular user, we also compared *average user* word annotation lists. *Average user word annotation lists* (either explicit or implicit) are created by taking the words from all individual lists (constructed as described above), and sorting by decreasing frequency (reflecting how many users have a word in their individual sets). After that, we select the top m most popular words, where m is the mean number of words that were chosen for explicit annotating across all users.

The average similarity scores for all answers between the average user word lists are displayed in the lower part of Table 3.5. The mean lists for each user group were created only from lists of users from that group. It can be seen that when the data from a greater number of users is aggregated, the more similar the final lists are (Jaccard coefficient between explicit and implicit mean lists is 0.361 for all users, while in divided equal users' parts it is only 0.309 and 0.328). The overlap for averaged words is on average higher than at the individual level. The agreement between explicit and implicit annotating is higher for correct answers. This could signify that *users are likely to understand explicitly which words indicate that an answer is correct. Conversely, their gaze is distributed more randomly when evaluating incorrect answers, suggesting that there aren't specific or consistent phrases that flag the incorrectness of an answer.* This is also supported by a higher number of areas of interest, and lower explicit and implicit overlap scores, between focus words for different users when dealing with incorrect answers.

3.4 Impact Of Highlighting Words

Highlighting terms was intended to assist users in evaluating answers. This section examines its impact by comparing the evaluations of two distinct groups: those who viewed answers with highlighted words and those who did not. We then analyze how closely the automated highlights align with the users' explicit and implicit annotations. Finally, we explore how the presence or absence of alignment between automated highlights and user annotations impacts the evaluation of answers, distinguishing between cases with high and low similarity.

Table 3.5: Similarity between explicit and implicit annotations.

Answer quality	User settings		
	all	without highlights	with highlights
<i>average of Jaccard coefficient for each user</i>			
ALL	0.284	0.321 [†]	0.247 [†]
INCORRECT	0.271*	0.306 ^{†*}	0.235 ^{†*}
CORRECT	0.299*	0.339 ^{†*}	0.259 ^{†*}
<i>Jaccard coefficient between "average user word list"</i>			
ALL	0.361	0.309	0.328
INCORRECT	0.337	0.304	0.301*
CORRECT	0.389	0.315	0.360*

* significant difference between correct/incorrect answer groups;

† between users shown/not shown automatic highlights

(Student's t-test, $p < 0.05$).

3.4.1 Assessment Quality

Users were randomly assigned to two groups with minimal differences in demographics, education, and English level. Group A included 7 females, 2 males, and 1 who preferred not to disclose gender, while Group B had 8 females and 2 males. Both groups consisted of students or university staff, with similar age distributions: Group A had 6 participants aged 18-30, 3 aged 31-40, and 1 over 40; Group B had 7 participants aged 18-30, 2 aged 31-40, and 1 over 40. Responses to the pre-task questionnaire — assessing interest, familiarity, and perceived difficulty — were consistent across groups. Thus, we can assume that *the main factor likely influencing their evaluation process was the presence or absence of automatically highlighted words in the answers*.

Table 3.6 shows that the overall accuracy of evaluation between the two groups (with or without automatic highlighting of words) is comparable. While users who saw highlighting evaluated high-quality correct answers more accurately (Student's t-test, $p < 0.05$), they sometimes mis-rated incorrect answers with low-quality correct answers. Regarding fine-grained answer quality aspects, the difference of ratings between the two groups was not statistically significant. *In other words, we can say that automatically highlighting words does not influence the accuracy of assessment of non-factoid answers.*

Table 3.6: Accuracy of rating corresponding to ground truth.

Answer quality	User Settings	
	without highlighting	with highlighting
ALL	0.72	0.71
INCORRECT	0.81	0.77
CORRECT	0.62	0.64
Low Quality & High Quality	0.58 & 0.67	0.54 & 0.8

3.4.2 Important Word Annotation

The three last columns of Table 3.3 show the extent to which people shown/not shown automatic highlighting agreed when marking a set of important words in an answer, both within and between the two user groups. There is good agreement between users within their groups; on average, users who were shown highlighting have slightly higher agreement to users who did not see it. The agreement was the lowest between users in two different groups, especially, for incorrect answers. Manually inspected cases showed that users who saw highlighted words in an incorrect answer had a tendency to annotate them as negative. It also explains higher agreement within this group for incorrect answers. At the same time, these assessors sometimes skipped annotating of those already highlighted words in correct answers, which is supported by the fact that the average count of annotated words in this group was lower than in the group without automatic highlighting, 7.8 versus 11.0.

In contrast, the agreement of the implicit vector of important words (the construction of which was described in the previous subsection) for users in the group with automatic highlighting was lower than in the group without highlighting (Table 3.4). This was a counter-intuitive result as we expected users would pay attention to the words which were specifically highlighted in the text. *It appears that users read highlighted words faster on average, and returned to them less frequently, studying other words instead.*

The agreement between explicit and implicit annotations in the group with automatic highlighting was also lower than for the group without. We can therefore conclude that, *despite the fact that automatic highlighting does not impact the accuracy of the answer evaluation, it can influence what parts of the answer users pay attention to during evaluation.* However, as we stated earlier, we could not find any significant differences in the correctness ratings between the groups.

3.4.3 Differences In Gaze Metrics

We compare gaze metrics between the two different user groups. As shown in Table 3.7, users from the groups with automatic word highlighting rated answer quality while spending less time on the answer screen. The difference between the two groups was significant, as was the difference for mean fixation durations and regression durations. While fixation and regression counts were not significantly different between the groups, mean fixation and regression durations and counts were lower for the group with highlighting, which could indicate that these users required less effort for answer evaluation. We also report all metrics for both groups with exclusion of all assessment errors. The trends show that users with automatic highlighted word rated an answer faster and with fewer fixations and regressions even when considering correct labeling only. *This finding supports the*

hypothesis that highlighting important words in an answer for a non-factoid question makes the evaluation process easier for a user.

Table 3.7: Gaze metrics for different answer quality levels.

Answer quality	Metric				
	Total view seconds	Mean fixation duration	Mean regression duration	Mean word fixations	Mean word regressions
<i>among all users</i>					
ALL	17.62	174.89	174.80	61.95	24.28
INCORRECT	16.89*	173.62	172.10*	60.22	22.83*
CORRECT	18.48*	176.37	177.93*	63.96	25.97*
<i>LQ</i>	17.39	176.50	179.75	60.61	25.31
<i>HQ</i>	20.08	176.17	175.26	68.87	26.93
<i>without guiding highlighting</i>					
ALL	18.66‡	181.80‡‡	181.14‡‡	63.12	25.44
INCORRECT	18.18‡	180.22‡‡	179.29‡‡	61.81	24.46
CORRECT	19.21	183.63‡‡	183.25‡‡	64.63	26.58
<i>LQ</i>	16.98	182.26‡‡	183.34‡	59.02	24.11
<i>HQ</i>	22.47‡	185.68‡‡	183.12‡‡	72.87	30.20
<i>with guiding highlighting</i>					
ALL	16.62‡	168.00‡‡	168.59‡‡	60.87	23.21
INCORRECT	15.59*‡	167.01‡‡	165.11*‡‡	58.63	21.20*
CORRECT	17.81*	169.15‡‡	172.67*‡‡	63.48	25.54*
<i>LQ</i>	17.80	170.74‡‡	176.21‡	62.20	26.52
<i>HQ</i>	17.83‡	166.80‡‡	167.40‡‡	65.36	24.09
<i>without guiding highlighting (users with 100% labeling accuracy)</i>					
ALL	18.31	181.15††	180.80††	63.91	26.19
INCORRECT	18.92	179.61††	178.29††	64.05	26.91
CORRECT	17.61	183.50††	184.54††	63.76	25.37
<i>LQ</i>	16.04	183.37†	187.88††	57.73	22.43
<i>HQ</i>	19.92	183.69††	180.33†	72.61	29.69
<i>with guiding highlighting (users with 100% labeling accuracy)</i>					
ALL	16.92	167.59††	165.93††	61.88	24.10
INCORRECT	16.13*	165.50*††	162.29*††	60.71	23.17*
CORRECT	17.82*	170.49*††	170.98*††	63.22	25.17*
<i>LQ</i>	18.47	171.65†	173.64††	62.81	27.26
<i>HQ</i>	16.92	169.35††	168.36†	63.79	22.24

* significant difference between correct/incorrect answer groups;

‡ difference between users shown/not shown highlighting;

† difference between 100% accurate users shown/not shown highlighting.

Student's t-test. One symbol $p < 0.05$, two symbols $p < 0.001$.

3.4.4 Similarity of Highlights and Annotations

To investigate the similarity of automatic highlighting (based on *TF-IDF* and capitalized words) and explicit or implicit user annotations, and how this influences evaluation, we first need to construct a *target user annotation* vector. As users who saw suggested highlighting are biased in their implicit and explicit word lists, we exclude such users from comparisons for the current analysis. Moreover, since the highlighting aims to help a user with evaluation, we only use the feedback of those users who correctly rated answers; otherwise we could end up with the misleading word suggestions. As an explicit or implicit user word annotation we used an "average user word list", which was described in Section 3.3.4. In the remainder of this analysis, we refer to these average user word lists, constructed only over data from users who did not see automatic word highlighting, as "*implicit target user annotating*" and "*explicit target user annotating*"

The first column of a top part of Table 3.8 shows the similarity calculated as the Jaccard coefficient between automatic word highlighting based on *TF-IDF* and explicit user annotations. The similarity is the highest for high quality answers which means it is mostly not capital or *TF-IDF* heavy words that allow people to understand the incorrectness of an answer. However, as we previously showed, the agreement on words among users is higher for correct documents as well. In the same way, we compared the shown highlighting with implicit user annotations. The results are reported in the first column of Table 3.8 on the bottom. Generally, the similarity is higher than with explicit annotation and this difference is statistically significant overall and among incorrect answers (Student's t-test, $t - \text{statistic} = 3.9$, $p < 0.01$). This could be explained by the fact that users could not always explicitly identify (and annotate) the complete list of important words which they

Table 3.8: Similarity between model and user annotations.

Answer quality	Suggestion type		
	shown TF-IDF	base BERT	fine-tuned BERT
<i>Jaccard coefficient with user explicit annotations</i>			
ALL	0.266	0.276	0.312
INCORRECT	0.274	0.256	0.269
CORRECT	0.261	0.299	0.363*
<i>LQ & HQ</i>	0.236 & 0.288	0.289 & 0.313	0.341 & 0.391
<i>Jaccard coefficient with user implicit annotations</i>			
ALL	0.295	0.314	0.316
INCORRECT	0.299	0.292	0.278
CORRECT	0.290	0.339	0.361*
<i>LQ & HQ</i>	0.274 & 0.314	0.347 & 0.328	0.372 & 0.346

* significantly different from *TF-IDF* (Student's t-test, $p < 0.05$).

implicitly paid attention to during the rating process. We observed this in the previous subsection as there was a low agreement between explicit and implicit word vectors for incorrect answers.

We further investigate how this similarity is connected with the user’s speed of evaluation. As we have shown, users from the group with automatic word highlighting evaluated answers statistically significantly faster than users who did not see highlighting. We found statistically significant weak negative correlation (Pearson -0.27 , $p = 0.02$) between the total view time of these users and the similarity of automatic word highlighting to the explicit user annotations. We also found low negative correlation of the view time with similarity to implicit words (Pearson -0.14 , $p < 0.001$). For completeness, we checked the correlation between the Jaccard scores and length of suggestions to reject the hypothesis that the quality of automatic word highlighting depends on their length; the correlation was not statistically significant (Pearson -0.05 , $p = 0.653$).

We also calculated the average similarity for the subset of answers where total view time results were inconsistent with the general trend (the outliers which were viewed longer by the users who were shown an answer with automatic word highlighting). Notably, their similarity with target explicit annotations for these answers is slightly lower with the value of 0.254, versus 0.275 for those answers that were rated more quickly by users who were shown automatic word highlighting. However, this difference is not statistically significant. From these observations, we can infer that *the more similar automatic word highlighting is to users’ annotations, the more helpful it is for them when evaluating answers*.

3.5 Analyzing Transformer Attention

In this section, we examine how the attention mechanisms of a Transformer model [187], specifically tuned for non-factoid question evaluation, align with the annotations made by users in our study, both explicit and implicit. We start by outlining the method used to extract key words from the attention maps of the Transformer model. We then provide details about the model itself, including its configuration and fine-tuning approach.

The core of this analysis involves comparing the words marked explicitly and implicitly by users in our study with the highlighting constructed from the Transformer model’s attention, as well as with our study’s baseline highlighting method. This comparison helps us understand whether the model’s focus areas are aligned with human attention and how this might enhance the evaluation of answers in non-factoid question answering systems.

3.5.1 Attention Construction

Self-attention assigns a weight (attention) from each word in a sentence to each other word, which can be interpreted as word importance and transformed into scores for a word highlighting algorithm. In our case, we input the query and answer simultaneously into a Transformer neural network, obtaining query-to-answer and answer-to-answer attention maps. Then, we calculate the importance of each answer token with respect to question tokens by averaging attention weights leading to this token from the question. We compute answer-to-answer token importance in the same fashion. Finally, we summed the aforementioned averages to form the attention score for each token:

$$\text{attention_score}(t) = \frac{\sum_{q \in Q} w_{qt}}{|Q|} + \frac{\sum_{a \in A, a \neq t} w_{at}}{|A|} \quad (3.2)$$

While our user study yielded data of a form appropriate for training a model, the quantity of labels is insufficient; we therefore use a pre-trained one, specifically a “large” uncased BERT model with whole word masking [51], which has 24 layers, each containing 16 attention heads. The BERT model was chosen due to its success in the similar task of a non-factoid answer ranking [130]. We study the average and standalone performance of attention maps of the last layer heads, as they have no shared parameters between them.

Given a BERT model, we constructed the input as **[CLS] <question> [SEP] <answer> [SEP]**, where **[CLS]** and **[SEP]** are special tokens that indicate the beginning of the input and the separation between the sentences, respectively. Using this form of input, we obtained attention maps for each question-answer pair. We excluded weights that were connected to the **[CLS]** and **[SEP]** special tokens, as they do not appear in user explicit or implicit annotation vector, and Clark et al. [40] demonstrated that these tokens have insignificant impact on the accuracy for most heads. We leveraged the BertVis visualization tool Vig [192] to extract attention maps from the model. Additionally, we wanted to infer the attention score for word-level tokens, but the BERT model uses byte-pair encoding as its tokenization method [168], leading to some words being split into sub-words. Thus, we had to deconstruct the BERT text representation by merging tokens starting with the special symbol ‘##’ and summing their attention scores to get an attention score for the original word, as was proposed by Clark et al. [40]. The final vector could be interpreted as an average importance weight for every word in an answer.

To extract more informative attention maps, we fine-tuned the base BERT model to the non-factoid QA evaluation task similar to the one users in our study were asked to do. Given a QA pair from the nfL6 dataset, the model had to predict if an answer was correct or not. First, we excluded QA pairs that were used in our user study since they also came

from this dataset. Then, we randomly chose 1% of the rest as holdout test data, and divided the remainder into training and validation data (80% and 20%, respectively). Then we fine-tuned the model for two epochs reaching 0.96 accuracy on the test set. After that, we discarded the last classification layer, and used the remaining layers to obtain attention maps for held out subset of QA pairs - the same as we used in our user study.

The word highlighting task can be formulated in a way similar to our user study task as annotating n important words from an answer for a non-factoid question. To construct automatic highlighting from BERT attention weights, we firstly need to predict n . We chose the average count of words that were explicitly annotated by users as the target number of highlighted words. However, such annotations may not be available for new texts. Thus, we predicted word annotation counts using linear regression, with answer length as the only feature (the correlation between the number of highlighted words and answer length is 0.7, Pearson, $p < 0.001$).

For a given answer, we therefore predicted n using regression, and then selected the top- n words in the answer with the highest attention scores. We used an average of attention maps from all heads of the Transformer layer as the overall BERT attention.

3.5.2 Comparison With User Annotations

To evaluate whether BERT and users in our study pay attention to the same words, we calculated the Jaccard similarity with explicit and implicit user annotations by the same way we did such comparison with shown TF-IDF highlighting in Section 3.4.4. The results are shown in Table 3.8. Overall, there is much better overlap with target user annotating comparing to the TF-IDF baseline. The difference between *TF-IDF* and fine-tuned BERT is statistically significant for correct answers ($p < 0.004$). Differences for all and incorrect answers are not statistically significant ($p = 0.065$ and $p = 0.83$, respectively). BERT attention similarity is higher compared to *TF-IDF* for correct answers, and lower for incorrect answers. It appears that weights for words signaling that the answer is incorrect could not be obtained from the average attention of all heads. We investigate this in the next subsection.

3.5.3 One Head Is All You Need

Previous research shows that in BERT models, only a small subset of heads is important for the target task [131, 194]. Inspired by this, we investigated whether we could use the attention map from a single “best” head on the last layer to obtain a final score that is at least not worse than from the average head approach reported earlier. Recall that the performance of our method is different for incorrect and correct documents; we therefore analyzed them

separately. Since the number of available documents is quite small (43 incorrect and 37 correct), we used leave-one-out cross-validation for the selection of the best head. For each document, we selected the best performing head on all documents except the current one. Then we calculated the similarity for that document using the selected head. The comparison between best-head performance and averaged attention score is reported in Table 3.9.

Notably, for incorrect and correct answers there was only one best head over all folds: head 13 for correct answers, and head 11 for incorrect answers. Six out of sixteen heads always perform worse than the average. The performance with head 11 and head 13, which give the best similarity with explicit highlighting, also give the second-best similarity with implicit annotations. The improved similarity for incorrect documents with head selection also becomes higher than baseline *TF-IDF* highlighting: 0.292 versus 0.274 for explicit incorrect, and 0.313 versus 0.299 for implicit incorrect (Student's t-test, $p < 0.05$). The overall improvement is promising and relatively stable across folds on average, indicating that, given explicitly annotated answers, we can select the best head and improve the method scores.

3.5.4 Implications of Highlighting

Our analysis suggest that automatic word highlighting significantly speeds up the process of evaluating answer correctness in the challenging task of non-factoid question answering tasks. Specifically, it leads to reducing total view time, fixation count, and regression count. Since the previously proposed highlighting method by Qu et al. [152] only moderately corresponds to the words that users find important and can sometimes lead to confusion, we suggest adopting a more advanced algorithm that leverages self-attention.

To improve over the baseline, we can consider the use of a more complex algorithm that makes use of self-attention - a key mechanism in state-of-the-art Transformer neural networks that allows them to track dependencies between words. As we have shown, highlighting based on the attention map of a fine-tuned BERT model is more similar to

Table 3.9: Best-head Jaccard coefficient for *fine-tuned* BERT.

Answers	With explicit		With implicit		
	best head	average	best explicit head	best head	average
INCORRECT	0.292 11 head	0.269	0.284 11 head	0.313 0 head	0.278
CORRECT	0.371 13 head	0.363	0.366 13 head	0.381 5 head	0.361

human gaze and explicit annotations and, hence, could be potentially used during the non-factoid QA evaluation process or as highlighting in answer snippets on a SERP.

3.6 Discussion

We now discuss the findings from the eye-tracking data analysis, comparison with transformer attention, and insights into user interaction with NFQA systems.

Our study confirms the importance of answer completeness in user evaluations, which supports previous research that highlights the need for comprehensive and detailed answers in NFQA [103, 125, 170]. Users require not only factual correctness but also thorough answers that cover multiple aspects of the question. This requirement highlights a potential limitation in using data from community question answering platforms like Yahoo!Answers, where responses, even those rated as “best” often fail to meet the expectations of users with limited topic knowledge. There may be other important criteria that need further research, which will be addressed in the next chapter.

Additionally, our study found that users’ interest in the question and perceived difficulty of the question also play a role in the evaluation, which aligns with findings from other research on user engagement and information retrieval [3, 82, 101]. These factors can influence the accuracy and reliability of user judgments. Given the popularity of long-form answers and the reliance on human evaluation for building gold test samples and benchmarking, there is a pressing need to develop methodologies to mitigate the complexity of evaluating unfamiliar, uninteresting, or difficult topics. In Chapter 6, we propose a framework aimed at improving the evaluation process, although it remains an open area of research.

Our findings indicate that users have a clearer understanding of what constitutes a correct answer, whereas their focus is more dispersed when evaluating incorrect answers. This aligns with the previous research by Qu et al., which suggests that users find it easier to recognize correct information than to pinpoint inaccuracies. We also support the hypothesis that visual aids can enhance the efficiency of the task [152] by demonstrating that answer presentation with highlighting significantly improves the time efficiency of the task. These findings align with previous research on the benefits of word highlighting for snippets in general [215]. However, the influence of highlighting on user attention patterns also suggests that it should be implemented carefully to avoid biases that might lead users to overlook important non-highlighted information.

Finally, our study adds a novel dimension to the research that explores the interpretability of transformer attention for various NLP tasks [23, 34, 186]. Our work, being the first to compare human gaze patterns with transformer attention, showed that the attention

mechanism of a fine-tuned BERT model closely aligns with user gaze patterns compared to simpler TF-IDF vectors, suggesting that transformer-based models can be used to identify key information in answers effectively. This alignment supports the efficiency of transformer-based models in enhancing the accuracy and efficiency of NFQA systems.

Overall, these findings highlight the complexities of user interaction with NFQA systems and provide valuable insights for further research on NFQA evaluation and presentation. The discussed results suggest that more work is needed in establishing robust evaluation criteria and frameworks to advance beyond current online metrics and improve NFQA system effectiveness.

3.7 Chapter Summary

In this chapter, we examined how users interact with passage-level answers to non-factoid questions, analyzing the influence of answer quality, users' prior knowledge, and how answers are presented. Through a comprehensive user study, we analyzed the words participants identified as important for assessing answer quality and the words that captured their attention, as indicated by eye-tracking data. This examination provided insights into how users evaluate information in non-factoid question answering scenarios.

We discovered that the completeness of an answer significantly affects its perceived correctness. Non-factoid answers should not only be factually correct but also thorough, covering all necessary aspects of the question such as explanations, diverse viewpoints, and examples. Additionally, our results indicated that a user's interest in the question and perceived difficulty impact the accuracy of their ratings, suggesting that these factors should be considered when evaluating answers to non-factoid questions.

Our study also showed that it is more challenging for users to identify incorrect parts of an answer than to confirm its correctness, as evidenced by higher agreement in annotations for correct answers both within and across users. This suggests that users have a clearer understanding of why an answer is correct, while incorrect answers lead to more dispersed user gaze and less overlap in annotations, indicating varied areas of interest and a scattered focus.

The analysis of highlighting of important words in answers demonstrated that they help users to rate answers more quickly and with less effort while maintaining the same level of quality. At the same time, highlighting changes implicit and explicit annotations from users, drawing their attention more to the words that are not highlighted. Furthermore, we found that the attention mechanism of a Transformer neural network, particularly using weights from a fine-tuned BERT model, aligns more closely with user gaze than baseline TF-IDF-based highlighting. This suggests that utilizing attention-based highlighting could

simplify the evaluation process in non-factoid QA.

The findings from this chapter, particularly highlighting the importance of answer completeness alongside correctness, lead us to our next area of investigation. We explore whether there are additional aspects that contribute to the usefulness of an answer in non-factoid question answering scenarios. Given that prior knowledge influences evaluation, we now aim to investigate questions genuinely of interest to a person and conduct an analysis from the asker’s perspective. The subsequent chapter examines these aspects in greater detail, presenting a taxonomy of 12 different qualities associated with answer usefulness. It also takes a step further toward creating more suitable metrics for NFQA by evaluating the effectiveness of commonly used answer quality metrics such as ROUGE, BLEU, METEOR [13], and BERTScore [217].

Chapter 4

Aspects of Human-Perceived Answer Utility in NFQA

In the previous chapter, we examined how people evaluate non-factoid answers using eye-tracking to understand where they focus and to compare this with neural network attention patterns. We discovered that the completeness of non-factoid answers, typically passage-length, is just as crucial as their correctness. This finding led us to explore further aspects that influence the overall utility of answers. In this chapter, we thoroughly examine a comprehensive list of aspects that define non-factoid question answering (NFQA) answer utility, introducing and analyzing a taxonomy of twelve different aspects that contribute to the value of an answer to the asker. This research enables us to identify the most important aspects of answer utility and to investigate the effectiveness of commonly used answer quality measures, including ROGUE, BLEU, METEOR, and BERTScore, within the context of non-factoid question answering (NFQA).

4.1 Introduction

In information retrieval, the quality of a document retrieved for a given query is usually assessed based on aspects such as the document's relevance to the query, its popularity, and the credibility of its source. However, in question answering, especially for non-factoid question answering (NFQA), these criteria alone may not suffice. Since non-factoid questions often require answers that provide detailed explanations or context, they pose unique challenges for assessing their usefulness.

This chapter focuses on identifying what makes an answer useful in non-factoid question answering. We explore the various aspects that contribute to answer utility, recognizing that these might differ significantly from those used for simpler, factoid queries. We began by developing a taxonomy of answer utility aspects from a set of sample questions and answers generated in a preliminary study. This taxonomy was refined over several iterations to include twelve key aspects that capture the utility of answers to non-factoid questions effectively.

Having created our aspect taxonomy, we conduct a multi-step editorial study, involving question/answer generation and usefulness/aspect assessment steps. In the question generation step, all participants act as askers and generate questions of different categories. In the answer generation step, we obtain answers to the generated questions from two different answer sources: participants, who act as answerers, and a commercial web search engine. Finally, in the usefulness/aspect assessment step, the askers assess the usefulness of the answers they received. Additionally, each answer is evaluated across the twelve aspects from proposed taxonomy.

Using the labels attained through this editorial study, we measure the correlation between the different answer aspects and the perceived usefulness of answers in order to answer the following research questions:

- What are the key aspects that make an answer useful to someone asking a non-factoid question?
- How do different factors related to the question and the person answering affect the perceived usefulness of answers?
- How do answers generated by humans compare with those generated by systems in terms of their utility?
- Are current metrics for evaluating answer quality adequate for capturing the utility of answers to non-factoid questions?

4.2 Taxonomy of Answer Utility Aspects

To begin our exploration of what makes an answer useful, we developed a taxonomy of aspects that could indicate answer usefulness. First, we created a draft taxonomy, involving more than a dozen potential aspects and extended by reviewing aspects used in other problem domains, such as online news [8] and community question answering (CQA) [170].

Secondly, a preliminary test study was performed to assess the relevance of each aspect in capturing the usefulness of answers. The study participants comprised two PhD students and one Senior Research Fellow, including one male and two females. Together, they evaluated 30 question/answer pairs specifically generated for this purpose. During this evaluation, the participants collaboratively determined the impact of each aspect on the perceived utility of the answers.

This process led to several modifications in the taxonomy. Aspects that did not significantly influence the perceived usefulness, such as those related to sentiment, were eliminated. Additionally, aspects regarding the source of the answers, like reputability and authoritativeness, were removed to simplify the focus of the study and to avoid the necessity of disclosing participant details in the subsequent study. This initial assessment led to the development of a draft taxonomy, which was then rigorously tested in a following controlled user study which involved a larger and more diverse group of participants who were not previously familiar with the taxonomy. This study is explained in detail in the next section. The final taxonomy includes the 12 aspects shown in Figure 4.1, where the aspects are grouped under four headings.

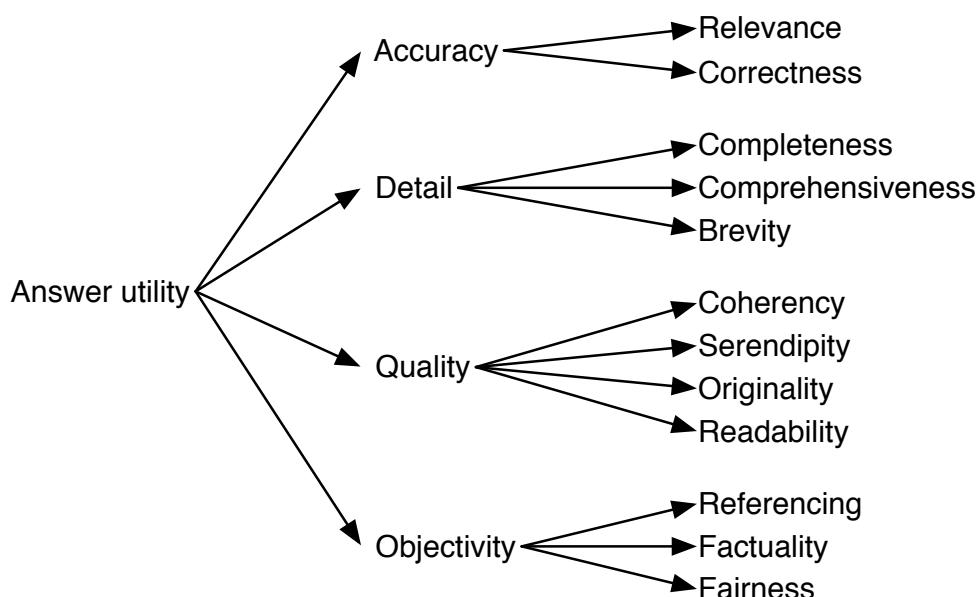


Figure 4.1: Aspects of answer utility.

Below is a brief description of each aspect included in the taxonomy:

1. The answer is *relevant* as it is about the subject of the question.
2. The answer is *correct* as it contains an accurate response to the question.
3. The answer is *complete* as it covers every aspect of the question.
4. The answer is *comprehensive* as it contains detailed information.
5. The answer is *brief* as it does not contain redundant information.
6. The answer is *coherent* as it does not contain inconsistent statements.
7. The answer is *serendipitous* as it contains some unexpected but positively surprising information.
8. The answer is *original* as it is not taken from another source.
9. The answer is *readable* as it is fluently written.
10. The answer is *referencing* additional information sources.
11. The answer is *factual* as it is based on things that are known to be true.
12. The answer is *fair* as it is free of any kind of bias.

We note that, although some aspects seem similar at first glance, they are semantically different. For example, comprehensiveness differs from completeness in that an answer may cover every aspect of the question but to different levels of detail. Similarly, factuality differs from correctness. Factuality refers to the information being based on verifiable evidence, while correctness means the answer is appropriate and accurate in context for a given user. For instance, take the question “Do aliens exist?”. There are various existing facts on this matter, based on which different answers could be constructed. These answers might be marked as correct by some users and incorrect by others, while both users will not debate over the truthfulness of the facts themselves. The answer could be “Aliens do not exist. No space agency has released evidence about their existence” or “Statistically, it is more probable that aliens do exist. The Drake Equation estimates there could be numerous habitable planets in the galaxy.”

4.3 Editorial Study

To thoroughly explore the process of non-factoid question answering, we conducted a controlled editorial user study aimed at achieving two main objectives: first, to assess the importance of various aspects of answer utility, and second, to explore how the structure and complexity of answer generation varies across different types of non-factoid questions.

This study involved twelve participants, recruited online through a university's participant database. Before the study began, participants were provided with an information sheet that detailed the study's objectives and procedures. Each participant signed a consent form to confirm their understanding of the study's purpose and their agreement to allow the use of their anonymized data for research. All communications with participants were conducted online via email.

Assessments were performed using spreadsheets distributed by the study coordinators. To gather comprehensive background information and feedback, participants completed pre-study and post-study questionnaires. Upon completion of all assigned tasks, which spanned multiple days with a total estimated time commitment of 20–25 hours, participants were compensated with a gift voucher valued at \$500 (AUD) each. The study was reviewed and approved by the Human Research Ethics Committee of RMIT University.¹

This chapter focuses on presenting the findings related to the various aspects of answer utility and their significance. The subsequent chapter introduces a newly developed taxonomy of non-factoid questions and their answers, detailing the differences in answer structures and the complexities of answer generation across various categories. This discussion includes all findings related to this taxonomy, which helps shape the research into non-factoid question answering.

4.3.1 Study Procedure

The editorial study consisted of three stages:

1. Question Generation and Labeling.
2. Answer Generation.
3. Answer Utility and Aspect Labeling.

In the first task, each participant acted as an asker and generated a number of questions that were used in the subsequent steps of the study. In the second task, each participant acted as an answerer and generated answers for questions received from an asker. Concurrent to

¹The IRB approval number is 23125.

this task, additional “system” answers were retrieved from the Google web search engine. In the third task, all answers were returned to the askers of respective questions, and then the askers labeled the various aspects of every answer they received for their questions, both human- and system-generated. The workflow of the editorial study is illustrated in Figure 4.2.

A preliminary trial was conducted prior to the official commencement of the study to ensure the effectiveness and clarity of the study procedures and the newly developed taxonomy. This trial involved two doctoral students and a postdoctoral researcher, who worked together to refine the study design. This phase included making necessary adjustments to the instructions and procedures, as well as developing preliminary expected answer structures for each question category.

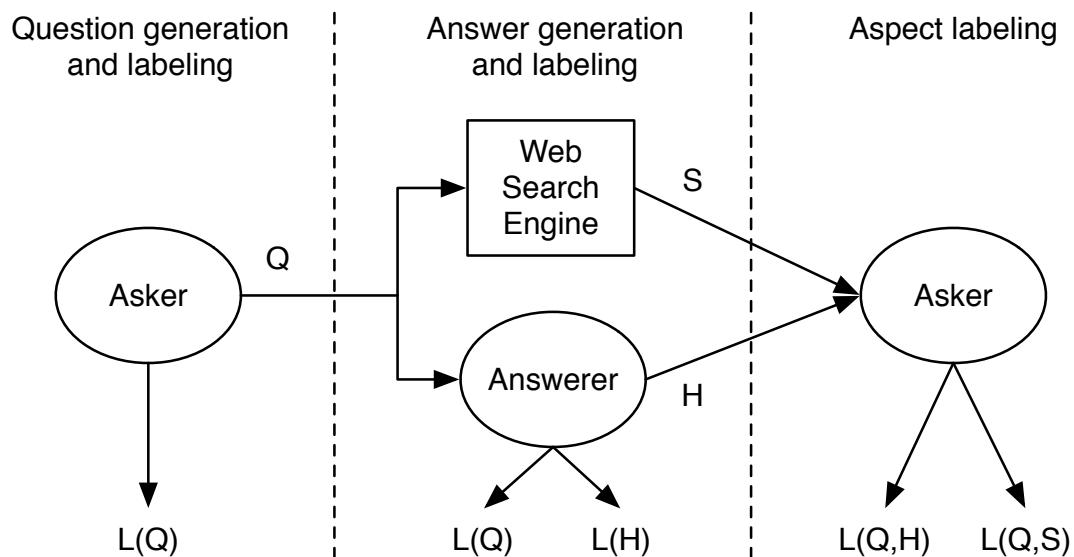


Figure 4.2: Editorial study workflow: Q is a question; H and S are human and system answers, respectively; and L are label sets obtained using Q, H, and S.

4.3.2 Pre-study Questionnaire

As an initial part of the editorial user study, participants were required to complete a pre-study questionnaire to collect demographic and background information:

- **Age:** Participant age ranged from 18 to 57, with an average of 32.
- **Gender:** There were more female participants ($f=10$, $m=2$); no participants identified as non-binary or chose ‘prefer not to say’. We do not expect this gender imbalance to affect our results substantially.

- **Language Proficiency:** Most participants (ten) were native English speakers, with the remaining two being proficient.
- **Educational Background:** Three participants had a master's degree, five had a bachelor's degree, one had a graduate certificate, and three had a high school degree.
- **Areas of Expertise:** Participants came from a diverse range of fields, including marketing, engineering, education, graphic design, communication, linguistics, financial planning, arts, economics, and accounting.
- **Computer Proficiency:** Ten participants reported using a computer multiple times daily, while two participants used a computer once a day.

4.3.3 Task 1: Question Generation and Labeling

The participants (askers) were first presented with a non-factoid question taxonomy containing six categories: INSTRUCTION (how to), REASON (cause/effect), EVIDENCE_BASED (description), COMPARISON, EXPERIENCE (advice), and DEBATE. A detailed discussion of this taxonomy, including its development, validation, and examples, as well as its applications in non-factoid question answering research, will be presented in the next chapter (refer to Table 5.1).

The categories were explained to the askers in detail, together with a sample set of question patterns for each question category. They were then asked to select four questions from their recent search history, i.e., queries they previously submitted to a web search engine, for each category. Additionally, they were asked to select four factoid questions, which typically expect short answer snippets (e.g., "Who is...", "When did...", "What is..."), to facilitate a comparative analysis. This resulted in each participant-asker generating a total of 28 questions.

Askers were required to provide well-structured questions that start with a question word and end with a question mark. The scope of the questions was not constrained, but the askers were instructed to avoid sensitive questions that could disclose their identity. If suitable, askers were allowed to convert keyword-based queries from their search history into full questions. They were also permitted to create new questions that genuinely interested them if no suitable queries were found in their search history. Out of the 336 questions generated, the majority, 63.1% were based on the askers' search histories, while the remaining 36.9% were newly created.

Generated questions were reviewed by the study coordinators. Approximately 12.8% of the questions were identified as problematic due to various issues: some were misclassified into incorrect question categories, others were poorly formulated, and one question was

a duplicate. Subsequently, another iteration was conducted with the askers to refine the non-factoid question categories and answer quality aspects within the taxonomies, and to correct or replace the problematic questions. The askers were then requested to evaluate each of their questions using a three-point scale to assess the following criteria:

1. **Question Importance:** How important is/was it for you to receive a useful answer to this question? The scale ranged from not important (0), somewhat important (1), to very important (2).
2. **Pre-search Question Difficulty:** How likely do you think it is for a human to provide a useful answer to this question without consulting any information sources? The options provided were not likely (0), somewhat likely (1), and very likely (2).
3. **Post-search Question Difficulty:** Considering a scenario where a human initially cannot provide a useful answer without additional information, how likely do you think it would be for them to provide a useful answer after consulting some information sources? The responses ranged from not likely (0), somewhat likely (1), to very likely (2).

The assessment of question difficulty was divided into two parts: pre-search and post-search. The pre-search difficulty was intended to gauge the inherent complexity and the knowledge required to answer the question independently. In contrast, the post-search difficulty aimed to evaluate the ease with which information sources could potentially aid in answering the question. To ensure an unbiased assessment, askers were prohibited from consulting any information sources during this phase of the study. Throughout the study, an information source was defined as an offline (e.g., a hard-copy book, a real-life friend) or an online (e.g., a web search engine, a community question answering site, a social media application) resource that could be used to seek a useful answer to a given question.

Table 4.1: Distribution of the labels from askers.

Criteria	Statistics		Labels		
	μ	σ^2	0	1	2
Question importance	0.80	0.73	0.38	0.43	0.19
Pre-search difficulty	0.80	0.65	0.33	0.54	0.13
Post-search difficulty	1.60	0.56	0.05	0.30	0.65

According to Table 4.1, 19% of the questions were deemed "very important" by their askers, while the proportion of questions deemed "not important" was roughly double this figure. The relatively low frequency of "very important" questions may be attributed to the

fact that many of these questions were not drawn from the askers' actual search histories but were instead generated during the study. When considering the source of the questions, which is not detailed in the table, we find that "very important" and "not important" questions each represent about one-fourth of the total questions from askers' search histories. On the other hand, "very important" questions are seven times less common than "not important" questions within the set of newly created questions.

As shown in Table 4.1, one-third of the questions are perceived (pre-search) as "not likely" to receive a useful answer, while only 13% are deemed "very likely" to obtain a useful response. The average pre-search likelihood score of receiving a useful answer is 0.80, suggesting that, on average, the generated questions tend to be perceived as difficult.

Contrastingly, the distribution of post-search assessments differs significantly. Post-search, askers believe that only 5% of the questions are "not likely" to receive a useful answer, with approximately two-thirds considered "very likely" to be answered effectively. This may be a sign of the participants' confidence in the wide-spread availability and easy accessibility of online information.

4.3.4 Task 2: Answer Generation and Labeling

Each set of 28 questions generated by an asker was provided to another participant (answerer), who was requested to come up at their disposal to formulate their responses. There were no constraints on the length of the answers; they could range from a few words to multiple paragraphs.

The main objective for the answerers was to provide responses that would be useful to the person who asked the respective questions. All responses were required to be in the answerers' own words. While answerers could paraphrase, summarize, or augment information from other sources with their own knowledge, direct copying and pasting from online sources were not allowed.

Prior to responding to each question, the answerers were required to provide the following information for every question:

1. **Answerer's interest:** How interested would you be in getting an answer to this question: not interested (0), somewhat interested (1), very interested (2)?
2. **Answerer's knowledge:** How likely are you to give a useful answer to this question without consulting any information source: not likely (0), somewhat likely (1), very likely (2)?

According to Table 4.2, the answerers were "very interested" in only a small portion of the questions they received (roughly, one out of every six questions). However, they were

Table 4.2: Distribution of the labels from answerers.

Criteria	Statistics		Labels		
	μ	σ^2	0	1	2
Answerer's interest	0.77	0.71	0.39	0.44	0.16
Answerer's knowledge	0.52	0.70	0.60	0.27	0.12

“moderately interested” or “very interested” in 60% of questions, an encouraging sign of participants’ engagement in the study. Regarding their knowledge about the question, the answerers felt that they were “not likely” to provide a useful answer for the majority of the questions (60%), while they claimed that they were “very likely” to provide a useful answer for only 12% of the questions. The latter value is close to the percentage of questions (13%) that were predicted by the askers as “very likely” to receive a useful answer (see the discussion associated with Table 4.1 in Section 4.3.3). However, when the two tables are compared, we also observe that the askers were more optimistic regarding the percentage of questions that were “not likely” to receive a useful answer (33% versus 60%).

After answering their questions, the answerers assessed the usefulness of their answers (“How useful do you think the answer you provided will be for a person who asks the corresponding question?”) using a 5-point scale: not useful (0), slightly useful (1), moderately useful (2), useful (3), very useful (4). Table 4.3 reports the distribution of labels for usefulness estimated by the answerers. We observe the distribution to be relatively balanced with the exception of the “not useful” label, which was selected for only 6% of the answers.

Table 4.3: Distribution of the labels assigned by the answerers regarding the estimated usefulness of their answers.

	Statistics		Labels				
	μ	σ^2	0	1	2	3	4
Usefulness	2.36	1.18	0.06	0.20	0.29	0.25	0.21

In addition to assessing the usefulness of their answers, answerers were required to identify the primary information source used to generate each response. The available options for sources included:

- Myself/another person
- Offline resource (e.g., a hard-copy book)
- Personal assistant (e.g., Alexa)
- Mobile or desktop application (e.g., Calendar)
- Search engine result page (e.g., Google)
- Community question answering website (e.g., Quora)
- News website (e.g., CNN)

- Other website
- Other resource

Answerers were specifically instructed to select the "search engine result page" option only if the answer was directly obtainable from the content displayed on the search engine's results page. They were cautioned against selecting this option if the search engine merely served as a gateway to another website that provided the actual answer.

Table 4.4: Sorted list of answer sources used by answerers.

Answer source	Frequency
Search engine result page	0.38
Myself (answerers themselves)	0.27
Other website (excluding CQA and news websites)	0.23
News website	0.04
Community question answering website	0.04
Another person	0.02
Mobile or desktop application	0.01
Offline resource	0.01
Personal assistant	0.00
Other resource (excluding any resource above)	0.00

Table 4.4 shows the frequency distribution of resources used by the answerers. The great majority of the answers are obtained from three main sources: a search engine result page; the answerers themselves; or other websites (excluding community question answering and news websites). Interestingly, no participant obtained their answers from a personal assistant, despite their increasing usage in homes and mobile phones. Also, no participant selected the “other resource” option, indicating that the provided categories sufficiently covered the range of resources typically used in question answering.

In addition to human answers, “system” answers were retrieved from Google web-search snippets for each of the 336 questions generated by the participants. The questions were submitted as queries, and then the answers were extracted from the snippets in the following priority, whichever was the first available:

1. **Featured Snippet Answer:** This appears at the top of Google’s Search Engine Results Page (SERP).
2. **Knowledge-Based Answer:** This appears on the right side of the SERP, often in a dedicated box.
3. **Top Result Snippet Answer:** This is extracted from the snippet of the most relevant search result and is always present.

Example snippets of each type are shown in Figure 4.3. When only the TOP RESULT SNIPPET ANSWER, which typically provides lower-quality answers, was available, it served as the last resort for extracting answers. If both the FEATURED SNIPPET ANSWER and the KNOWLEDGE-BASED ANSWER modules were available, the FEATURED SNIPPET ANSWER was prioritized due to its typically higher relevance and quality. This way, we obtained a single system answer for each question in our sample. All answers were scraped manually to prevent erroneous query rewriting, occasionally performed by Google.

In total, 57% system answers came from FEATURED SNIPPET ANSWER, 32% from TOP RESULT SNIPPET ANSWER, and 11% from KNOWLEDGE-BASED ANSWER.

Figure 4.3: Three types of Google SERP snippets available for the question “What is harry potter about?”

The screenshot shows a Google search results page for the query "What is harry potter about?".

- Search Bar:** What is harry potter about?
- Navigation:** All Images Videos News Shopping Books Maps More Tools
- Section 1 (Featured Snippet Answer):**
 - Title:** Harry Potter
 - Subtext:** Novel series
 - Buttons:** Overview, Books, Characters, Videos
 - Image:** A collage of five Harry Potter characters (Harry, Ron, Hermione, etc.)
 - Text:** The series follows the life of a boy named Harry Potter. In the first book, Harry Potter and the Philosopher's Stone, Harry lives in a cupboard under the stairs in the house of the Dursleys, his aunt, uncle and cousin, who all treat him poorly. At the age of 11, Harry discovers that he is a wizard.
- Section 2 (Top Result Snippet Answer):**
 - Source:** Wikipedia
 - Text:** Harry Potter - Wikipedia
 - Feedback:** About featured snippets • Feedback
- Section 3 (Knowledge-Based Answer):**
 - Section Header:** About
 - Text:** Harry Potter is a series of seven fantasy novels written by British author J. K. Rowling. The novels chronicle the lives of a young wizard, Harry Potter, and his friends Hermione Granger and Ron Weasley, all of whom are students at Hogwarts School of Witchcraft and Wizardry. [Wikipedia](#)
 - Character:** Hermione Granger, Harry Potter, Lord Voldemort • More
 - Genre:** Fantasy
 - Published:** 26 June 1997 – 21 July 2007
 - Feedback:** Feedback
- Related Content:** People also search for Harry Potter books (with thumbnail images).

4.3.5 Task 3: Answer Utility and Aspect Labeling

In this final task, the askers were provided with two different answers (a human answer and a system answer) for each of the 28 questions they have previously generated, forming 56 question–answer pairs in total. The askers were then requested to assess the usefulness of the answers they received (“How useful do you think the provided answer is from your

point of view?”) using a 5-point scale: not useful (0), slightly useful (1), moderately useful (2), useful (3), very useful (4). During the assessment of usefulness, they were allowed to consult any information source, except for Google, since the system answers they received had been retrieved from Google.

Table 4.5 shows the distribution of the usefulness ratings assigned by the askers to the answers they received. According to the mean values reported in the table, the askers seem to find human answers slightly more useful than the system answers. However, it is important to refrain from drawing premature conclusions based on this observation. As detailed later in Section 4.4.3 and the subsequent chapter, the perceived usefulness of system answers varies significantly across different question categories and is also influenced by which system module the answer was retrieved from.

Table 4.5: Distribution of the perceived usefulness labels assigned by the askers for human and system answers.

	Statistics		Labels				
	μ	σ^2	0	1	2	3	4
Human	2.47	1.25	0.10	0.14	0.20	0.33	0.24
System	2.33	1.47	0.18	0.15	0.14	0.25	0.29

Finally, the askers assessed the aspects for each of the answers they received. They were presented with the 12 statements given in Section 4.2. They expressed their level of agreement using a 5-point Likert scale:

- Strongly disagree (0)
- Disagree (1)
- Neither agree nor disagree (2)
- Agree (3)
- Strongly agree (4)

The order of questions was randomized before they were assessed. Participants were also encouraged to take a short break after completing the evaluation of each aspect to ensure attentiveness and reduce fatigue.

4.3.6 Post-study Questionnaire

The participants completed a post-study questionnaire to provide feedback on their overall experience with the study, their perception of its difficulty, and their views on the fairness of the compensation provided. The results were as follows:

- **Overall Experience:** Ten participants expressed satisfaction with the study, while two were neutral. None reported dissatisfaction.

- **Study Difficulty:** Three participants found the study easy, seven rated it as medium difficulty, and two found it challenging.
- **Compensation Fairness:** The majority, ten participants, felt the compensation was fair. Two participants believed the compensation should have been less, and none thought it should be more.

These responses indicate that the majority of participants were satisfied with their study experience, found the difficulty level to be manageable, and considered the compensation to be fair.

4.4 Answer Utility Aspect Analysis

In this section, we analyze the outcomes from the editorial study to deepen our understanding of how non-factoid question answering process. Using Spearman’s rank correlation [136] suitable for our ordinal data, we assess the correlations between various aspects of answer utility. The result tables indicate statistical significance for a standard test of the null hypothesis that the correlation is zero. We apply a Bonferroni correction to the p -values to account for multiple hypothesis tests.

For comparisons between different correlation coefficients, we follow the method described by Myers and Sirois [137], who use Fisher’s z-transformation for correlation coefficients and then test the null hypothesis that $\rho_1 - \rho_2 = 0$. When referring to the strength of the correlation values reported in the tables, we adopt the scale and terminology used by Prion and Haerling [151]: [0.00, 0.20) (negligible), [0.20, 0.40) (weak), [0.40, 0.60) (moderate), [0.60, 0.80) (strong), and [0.80, 1.00] (very strong). The labels obtained from the answerers and askers regarding the usefulness of answers are referred to as the estimated and perceived usefulness, respectively.

4.4.1 Key Aspects of Perceived Usefulness

To analyze the impact of various aspects on the perceived usefulness of answers, we display the mean values of each aspect for different levels of usefulness, using radar graphs in Figure 4.4. Additionally, we calculate correlation values between these aspects and perceived usefulness. The computed values are reported in the right-most columns of Tables 4.6 for non-factoid questions (on the top) and factoid questions (on the bottom), together with the correlation values between pairs of aspects. When discussing the results in this section, we focus only on statistically significant correlations and their significant differences (Bonferroni corrected equivalent of $\alpha = 0.05$ and $\alpha = 0.001$, see table captions).

Table 4.6: Correlations between the perceived usefulness and aspects of answers (non-factoid and factoid questions).

	Aspects (Non-Factoid Questions)										Perceived usefulness	
	Relevance	Comprehensiv.	Correctness	Coherency	Factuality	Readability	Brevity	Fairness	Serendipity	Referencing	Originality	
Completeness	0.67**	0.66**	0.65**	0.52**	0.42**	0.50**	0.41**	0.27**	0.09	0.11	0.65**	
Relevance		0.49**	0.59**	0.50**	0.44**	0.35**	0.47**	0.34**	0.32**	-0.02	0.15*	0.65**
Comprehens.			0.55**	0.45**	0.46**	0.40**	0.34**	0.41**	0.29**	0.21**	0.01	0.63**
Correctness				0.55**	0.58**	0.40**	0.45**	0.47**	0.26**	-0.04	0.01	0.60**
Coherency					0.46**	0.62**	0.52**	0.43**	0.33**	0.03	0.13	0.51**
Factuality						0.43**	0.33**	0.57**	0.13	0.02	-0.03	0.47**
Readability							0.39**	0.39**	0.23**	0.04	0.13	0.37**
Brevity								0.38**	0.13	0.10	0.14*	0.37**
Fairness									0.14*	-0.07	-0.07	0.33**
Serendipity										0.07	0.14	0.26**
Referencing											0.33**	0.08
Originality												0.08

	Aspects (Factoid Questions)										Perceived usefulness	
	Relevance	Comprehensiv.	Correctness	Coherency	Factuality	Readability	Brevity	Fairness	Serendipity	Referencing	Originality	
Relevance	0.65**	0.69**	0.44**	0.18	0.42*	0.45**	0.28	0.27	0.32	0.47**	-0.16	0.55**
Correctness		0.61**	0.48**	0.18	0.47**	0.44**	0.25	0.20	0.35*	0.41*	-0.16	0.53**
Completeness			0.42*	0.18	0.45**	0.48**	0.46**	0.08	0.36*	0.26	-0.19	0.52**
Factuality				0.11	0.56**	0.57**	0.46**	0.06	0.43**	0.43*	-0.01	0.43*
Serendipity					0.11	0.04	0.32	0.23	0.03	-0.18	0.27	0.36*
Coherency						0.55**	0.29	0.06	0.73**	0.42*	-0.23	0.34
Fairness							0.20	0.20	0.39*	0.55**	-0.19	0.31
Comprehens.								-0.16	0.24	0.03	0.23	0.28
Originality									-0.14	0.20	0.25	0.23
Readability										0.30	-0.16	0.17
Brevity											-0.08	0.13
Referencing												-0.01

* significance level $p < 0.000641$, ** significance level $p < 1.282e-05$. (Bonferroni corrected from $p = 0.05$ and $p = 0.001$, 78 tests).

For non-factoid questions, the aspects of *relevance*, *completeness*, *correctness*, and *comprehensiveness* show strong correlations with perceived usefulness. This finding is consistent with insights from the previous chapter, where *completeness* was found to be as important as *correctness*. Coherency and factuality also have moderate correlations with perceived usefulness. Besides, the correlations of these four aspects with usefulness are statistically significantly higher ($p < 0.00076$, Bonferroni corrected from $p = 0.05$, 66 tests) than the correlation between any other aspect, except for coherency and factuality. Out of 66 possible pairs of aspects, only four pairs show strong correlations, indicating that our taxonomy captures a diverse and semantically distinct set of aspects. Three of these four pairs include aspects strongly correlated with usefulness. The moderate correlation between correctness and factuality supports our earlier differentiation of these two aspects (see Section 4.2). Finally, the mean values displayed in Figure 4.4a indicate that answers to non-factoid questions are generally perceived as useful even when they are not original and not supported by references, as long as some influential aspects are present. In the same figure, we also observe that “not useful” answers are more likely to be incomplete or less detailed than being irrelevant or incorrect.

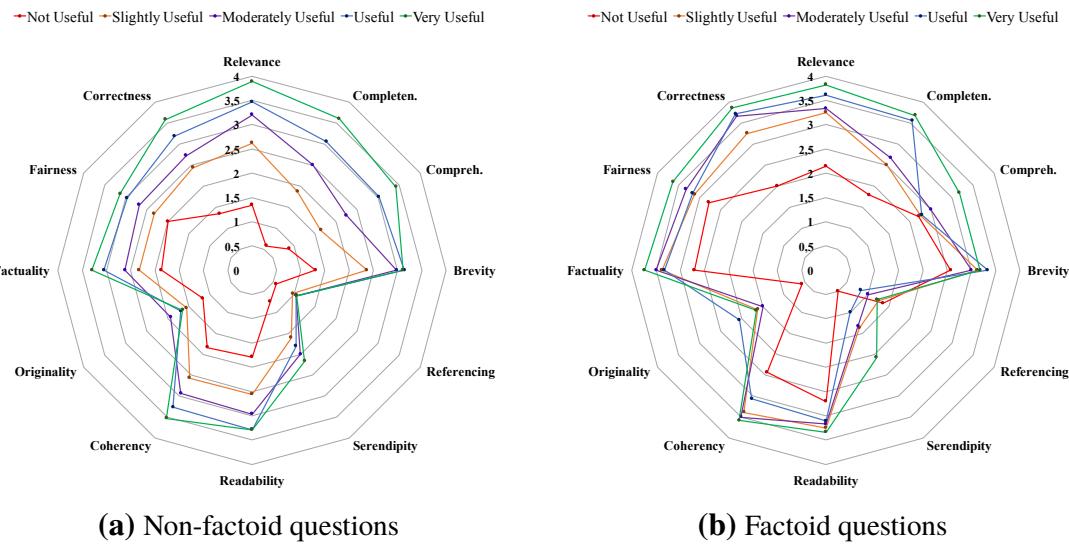


Figure 4.4: Mean values of different aspect labels.

For factoid questions, none of the aspects have a strong correlation with usefulness. The aspects of relevance, correctness, completeness, and factuality show only moderate correlations. Comprehensiveness, which is an important aspect for non-factoid questions, is absent in this list, probably because most factoid questions are likely to be answered satisfactorily with a phrase or a short sentence. In our data, “useful” and “very useful” answers have a mean length of 134.1 and 252.2 characters for factoid and non-factoid questions, respectively (Student’s t -test, $t = -5.68$, $p < 0.001$), supporting this claim.

4.4.2 Factors Influencing the Perceived Usefulness of Answers

Following observations from the previous chapter where certain factors like assessor interest and question difficulty influenced the evaluation process of NFQA, this section explores how similar factors affect an individual's ability to produce a useful answer. These factors include the *importance of the question to the asker, the question's difficulty both before and after searching for information, the answerer's interest in the question, the answerer's knowledge on the topic, and the answerer's self-assessment of the usefulness of their responses.*

Intuitively, one may expect that less difficult questions or questions which the answerer is interested in or has knowledge about are easier to answer, and thus, they should receive more useful answers. However, the correlation analysis detailed in Table 4.7 shows that all

Table 4.7: Correlation of attributes with perceived usefulness of answers.

	Estimated Usefulness	Perceived Usefulness
Importance	0.05	-0.05
Pre-search diff.	0.04	0.06
Post-search diff.	0.24**	0.14
Interest	0.18	-0.02
Knowledge	0.04	0.06
Est. usefulness	1.00**	0.16

** significance level $p < 1.19e-05$. (Bonferroni corrected from $p = 0.05$, 12 tests).

Table 4.8: Correlation of attributes with answer aspects.

	Aspects (Part 1: Relevance to Referencing)					
	Relevance	Correctness	Completeness	Comprehensiv.	Brevity	Referencing
Importance	-0.11	-0.17	-0.14	0.00	-0.04	0.02
Pre-search diff.	0.12	0.10	0.10	0.12	0.06	0.02
Post-search diff.	0.08	0.33**	0.13	0.11	0.19*	-0.12
Interest	-0.06	0.00	0.05	0.12	-0.12	-0.05
Knowledge	0.02	-0.02	0.11	0.20*	-0.05	0.12
Est. usefulness	0.27**	0.29**	0.30**	0.12	0.12	-0.22*

	Aspects (Part 2: Serendipity to Fairness)					
	Serendipity	Readability	Coherency	Originality	Factuality	Fairness
Importance	-0.05	-0.01	-0.06	0.00	-0.07	0.01
Pre-search diff.	0.01	0.02	0.11	0.06	0.05	0.05
Post-search diff.	-0.20*	0.01	0.12	-0.18	0.24**	0.31**
Interest	-0.03	-0.05	-0.13	-0.06	0.02	-0.04
Knowledge	-0.10	-0.03	-0.08	0.01	0.09	0.01
Est. usefulness	-0.04	0.14	0.13	-0.18	0.32**	0.33**

* significance level $p < 0.000595$, ** significance level $p < 1.19e-05$. (Bonferroni corrected from $p = 0.05$ and $p = 0.001$, 72 tests).

these attributes have negligible correlations with perceived usefulness (although none of the reported results are statistically significant). This may be attributed to the answerers' ability to access extensive information sources, thus compensating for any initial lack of knowledge or interest. For example, an answerer with limited knowledge about a challenging question might still manage to provide a useful answer after consulting various information sources, with 112 out of 203 such questions being rated as useful.

The aspect correlations shown in Table 4.8 provide deeper insights into how factors such as correctness, completeness, factuality, and fairness relate to influencing attributes like question difficulty, importance, and knowledge. The minimal correlations observed may be explained by the answerers' ability to effectively use diverse information sources, compensating for initial gaps in knowledge or interest. For instance, even when answerers lacked prior knowledge on challenging questions, they were still able to deliver useful answers after consulting information sources—112 out of 203 such questions were rated as useful.

Notably, there is negligible correlation between the usefulness of answers as estimated by the answerers themselves and as perceived by the askers. This finding suggests that *answerers often struggle to accurately judge what constitutes a useful answer from the perspective of the asker*.

4.4.3 Comparison of Human and System-Generated Answers

This section explores the differences in perceived usefulness between human and system-generated answers, focusing on the distinction between factoid and non-factoid questions. We further categorize the responses based on the source module on Google's Search Engine Results Page (SERP): Featured Snippet Answer, Knowledge-Base Answer, and Top Result Snippet Answer. In Table 4.9, we report the mean and standard deviation for the perceived usefulness of human and system answers, separately, for each question group. We also show the difference between the mean values for perceived usefulness of human and system answers in the last column. In the subsequent chapter, we will study the differences in perceived usefulness between human and system-generated answers more thoroughly, focusing specifically on non-factoid questions and comparing different categories within them.

In Table 4.9, the reported differences for knowledge-based answers do not show statistical significance for factoid or non-factoid questions, potentially due to small sample sizes (reported in the seventh column of the table). Hence, we do not analyze and discuss these results further. Focusing on non-factoid questions only, we observe that the featured snippet

Table 4.9: Various statistics about the perceived usefulness of human and system answers.

Question Type	Answer type	Human		System		#	$\mu_h - \mu_s$
		μ_h	σ_h^2	μ_s	σ_s^2		
All	Featured Snippet Answer	2.43	1.25	2.88	1.19	192	-0.44*
	Knowledge-Based Answer	2.53	1.41	3.08	1.15	38	-0.55
	Top Result Snippet Answer	2.53	1.21	1.07	1.21	106	1.46*
	All	2.47	1.25	2.33	1.47	336	0.15
Factoid	Featured Snippet Answer	2.38	1.47	3.05	1.28	21	-0.67
	Knowledge-Based Answer	2.74	1.63	3.53	1.07	19	-0.79
	Top Result Snippet Answer	3.38	1.41	1.75	1.98	8	1.63
	All	2.69	1.53	3.02 †	1.45	48	-0.33
Non-Factoid	Featured Snippet Answer	2.44	1.23	2.85	1.18	171	-0.42*
	Knowledge-Based Answer	2.32	1.16	2.63	1.07	19	-0.32
	Top Result Snippet Answer	2.46	1.17	1.01	1.13	98	1.45*
	All	2.44	1.20	2.21†	1.44	288	0.23

* sig. diff. btw. human/system $p < 0.05$ (Tukey HSD).

† sig. diff. btw. factoid/non-factoid $p < 0.05$ (Tukey HSD).

answers are perceived as more useful than human answers on average (the difference of means is -0.42). On the contrary, the top result snippet answers answers are perceived as somewhat less useful than human answers (the difference of means is 1.45). When we compare the mean values for factoid and non-factoid question sets, we see that the system answers given to factoid questions are perceived as statistically significantly more useful (3.02 versus 2.21), while there is no statistically significant difference in the case of human answers (2.69 versus 2.44). This may imply that modern search engines are effective in answering factoid questions, but non-factoid questions have room for improvement.

Figure 4.5 compares the mean values of aspects for human and system answers. Since the knowledge-based answers have a small sample size and have similar quality to the featured snippet answers, we combine the two sets in our analysis (Figures 4.5a and 4.5b). We display the mean values for top result snippet answers only for non-factoid questions (Figure 4.5c) since we have just eight factoid questions with a top result snippet answer.

4.4.4 Answer Quality Metrics Evaluation

Finally, we assess the suitability of metrics commonly used for evaluating the quality of non-factoid question answering systems — ROUGE [124], BLEU [144], METEOR [13], and BERTScore [217]. These measures yield a quality score for a given question, candidate answer pair by comparing the candidate answer with a reference answer which is assumed to be a perfect answer for the question. Herein, we aim to understand the effectiveness of

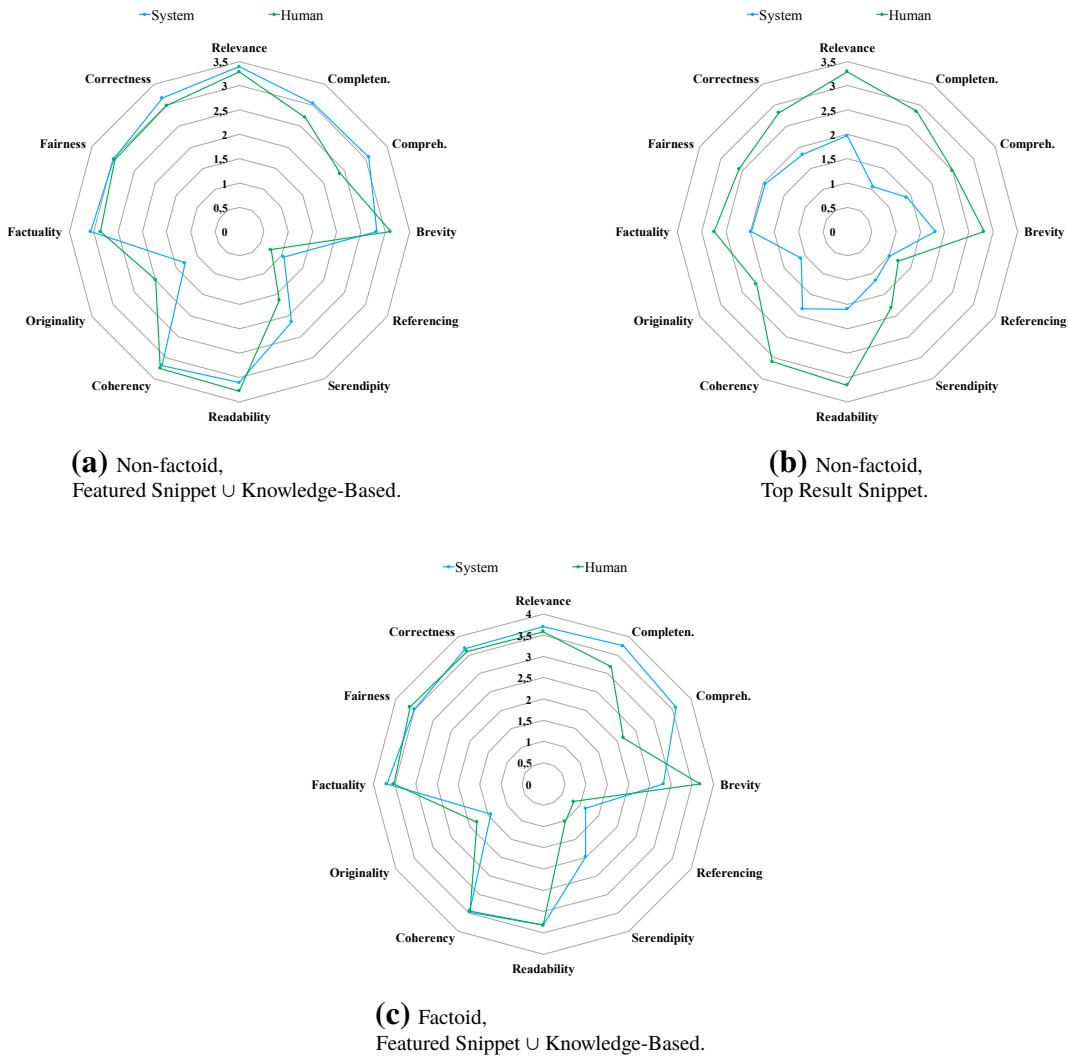


Figure 4.5: Mean values of aspect labels for human and system answers.

these measures in capturing the perceived usefulness of answers as determined by askers, specifically focusing on non-factoid questions.

To obtain a sample set of questions with reference and candidate answers, we adopt the following process: We first remove questions with no answers labeled as “useful” or “very useful”, leaving 216 questions (out of the 288 questions available). For each of the remaining questions, we select one of the two answers (a human and a system answer) associated with the question as a reference answer. Here, we prefer “very useful” answers over “useful” answers. The ties (about 20% of the remaining questions) are broken by selecting the answer with the higher mean value computed over all aspects. Of the 216 reference answers obtained, 116 are human answers, and the remaining 100 are system answers. The answers that are not selected in the process are used as candidate answers. The reference answers are statistically significantly more useful than the candidate answers,

on average: 3.51 versus 1.92 (Student’s t -test, $t = 17.6$, $p < 0.01$).

We compute the four measures above using our reference and candidate answers. Table 4.10 shows the correlation between the computed measures and the perceived usefulness of answers. Although all metrics show statistically significant correlations with perceived usefulness, these were generally weak to moderate, with BERTScore slightly outperforming the others. Additionally, these metrics correlated well with key aspects previously identified as crucial—relevance, correctness, completeness, and comprehensiveness. However, no metric displayed a statistically significant difference compared to others in terms of their correlations with these aspects. Notably, the highest correlation observed with BERTScore remains only moderate, suggesting that while these metrics provide some insights, there is a pressing need to develop more tailored metrics specifically for non-factoid question answering.

Table 4.10: Correlations of quality measures with the perceived usefulness of answers and their aspects.

	Aspects (Part 1: Relevance to Referencing)					
	Relevance	Correctness	Completeness	Comprehensiv.	Brevity	Referencing
ROUGE-L	0.36**	0.41**	0.31**	0.27*	0.35**	-0.11
ROUGE-1	0.37**	0.42**	0.31**	0.30**	0.33**	-0.12
ROUGE-2	0.33**	0.44**	0.28*	0.28*	0.35**	-0.06
BLEU	0.36**	0.43**	0.33**	0.37**	0.26*	-0.07
METEOR	0.35**	0.40**	0.28*	0.23*	0.40**	-0.09
BERTScore	0.33**	0.42**	0.36**	0.32**	0.26*	-0.13

	Aspects (Part 2: Serendipity to Perceived Usefulness)					
	Serendipity	Readability	Coherency	Originality	Factuality	Perceived Usefulness
ROUGE-L	0.19	0.25*	0.29*	-0.03	0.28*	0.23*
ROUGE-1	0.22	0.26*	0.32**	-0.08	0.26*	0.23*
ROUGE-2	0.19	0.22	0.28*	-0.08	0.28*	0.25*
BLEU	0.27*	0.19	0.23*	-0.16	0.29**	0.24*
METEOR	0.15	0.27*	0.31**	0.03	0.25*	0.28*
BERTScore	0.19	0.25*	0.33**	-0.06	0.31**	0.22
						0.42**

* significance level $p < 0.000641$, ** significance level $p < 1.282e-05$. (Bonferroni corrected from $p = 0.05$ and $p = 0.001$, 78 tests).

4.5 Chapter Summary

This chapter has investigated the impact of various aspects on the perceived usefulness of answers in non-factoid question answering scenarios. Building on findings from the previous chapter that completeness is as critical as correctness for non-factoid answers, we introduced and validated a taxonomy of twelve distinct aspects of answer utility. Our analysis confirmed that four aspects—relevance, correctness, completeness, and comprehen-

siveness—are strongly correlated with perceived usefulness, underscoring that non-factoid answers must be not only correct but also complete, relevant, and comprehensive. This indicates that when evaluating usefulness, users prioritize the accuracy and detail of an answer over its stylistic quality or subjective elements.

In the previous chapter, we discussed how an annotator’s interest in a question and their expertise influenced their ability to evaluate the quality of non-factoid answers when given no access to additional resources. In contrast, this chapter presents a different scenario: here, individuals who generated the questions (askers) had their questions answered by others (answerers) who had access to various information sources. This setup demonstrates that the factors such as an answerer’s interest in the question and their expertise have much less impact, we observed negligible correlations with the perceived usefulness, on their ability to generate useful answers when they can consult external resources. Consequently, this suggests that for askers, the perceived usefulness of an answer is primarily influenced by the content of the answer itself, rather than by the personal attributes or conditions of those generating the answers. This aligns with the work of Hart and Sarma [78], who highlighted the limited role of social cues in the perception of information utility.

Additionally, this chapter compared the perceived usefulness of answers provided by humans and automated systems. Results indicate that system-generated answers, particularly those from Google’s featured snippet and knowledge-base modules, are often seen as more useful than those provided by humans. This finding is consistent with advancements in AI-driven question answering systems which, in some cases, have demonstrated capabilities surpassing human performance [141, 155, 205]. It is important to note that these advanced system answers were primarily available for factoid questions or certain non-factoid answer categories. The next chapter will delve deeper into the performance of the systems across various non-factoid question categories including more challenging ones.

Furthermore, our evaluation of common answer quality metrics — ROUGE, BLEU, METEOR, and BERTScore — revealed that while BERTScore slightly outperformed the others, all metrics showed only weak to moderate correlations with the most critical aspects of answer usefulness. This suggests a need for developing more effective metrics tailored specifically for non-factoid question answering.

The findings of this chapter carry significant implications for the design and optimization of question answering and search systems. Most systems currently focus on the relevance and correctness of answers, but our results advocate for equal emphasis on completeness and comprehensiveness. This poses a challenge, particularly for web search engines, which must balance the detail and length of answers against the constraints of display space. Additionally, the potential benefits of enriching factoid answers with supplemental information suggest avenues for enhancing user satisfaction and engagement. Lastly, the

marked difference in user perceptions of standard search results and more sophisticated answer formats like featured snippets highlights important considerations for the trade-offs involved in snippet generation for search engines.

This chapter has identified key aspects that affect the perceived usefulness of answers in non-factoid question answering and highlighted the need for developing tailored metrics for a more accurate evaluation. We've explored how aspects like relevance, correctness, completeness, and comprehensiveness play significant roles. However, one critical aspect yet to be addressed is the variation in answer structures and evaluation processes for different types of non-factoid questions—from straightforward descriptions to complex queries requiring synthesis of multiple viewpoints. In the next chapter, we will explore the differences among various non-factoid question types, introducing and detailing the first comprehensive taxonomy of non-factoid questions and their expected answers.

Chapter 5

A Non-Factoid Question-Answering Taxonomy

In Chapter 4, we explored what makes an answer to a non-factoid question useful and compared these insights with those for factoid questions. We pointed out the need for more detailed metrics to effectively gauge their usefulness. However, non-factoid questions themselves can vary widely in complexity and in the types of answers they require. Until now, our focus has been on passage-length responses for non-factoid answers, but it became clear that different categories may need distinct answer structures. For example, some answers might be best expressed through a passage containing a list of facts, while others, like how-to questions, might require step-by-step guidelines.

This observation led us to investigate the distinct categories of non-factoid questions, the specific forms of answers they require, and whether current evaluation benchmarks adequately cover these variations. In this chapter, we introduce the first streamlined taxonomy of non-factoid questions and their corresponding answers, validated through extensive crowdsourcing and user studies. We will examine the differences in complexity across these categories, assess whether existing benchmarks provide a balanced representation, and evaluate the capability of current systems to address the diverse needs of non-factoid question answering.

5.1 Introduction

State-of-the-art systems significantly lag behind human performance in non-factoid question answering (NFQA), as demonstrated by datasets like NFL6, ANTIQUE, NLQuAD, and ELI5 [43, 59, 79, 173] [59, 106, 173]. Additionally, industry QA systems, whether generating snippets on search engine results pages (SERPs) or functioning within conversational agents, struggle to provide meaningful answers to complex non-factoid questions, such as *“If scientifically possible, should humans become immortal?”*. Such questions require answers that are not only correct but also nuanced and enriched with diverse viewpoints.

Despite advancements, there is no guarantee that systems will generalize across all non-factoid question categories. In fact, no analysis has been conducted on the distribution of question categories in NFQA datasets, beyond considering starting question words. This oversight risks ignoring under-represented categories, focusing instead on popular and simpler cases.

The ultimate goal of NFQA could be an end-to-end system that can handle all categories of non-factoid questions. At this stage, however, it is beneficial to study each question category separately to address their unique challenges. For example, the question *“How to come up with ideas?”* requires a descriptive answer with concrete steps, whereas *“What is the meaning of nkg?”* needs a list of definitions with explanations and examples. Understanding the specific answer structures required for each category is essential for advancing NFQA. To do so, we must first define possible question categories.

Currently, there is no unified and well-evaluated taxonomy for NFQs, unlike factoid QA where a few taxonomies of question categories and forms of target answers exist [87, 123, 180]. While some related works (described in Section 2.5.3) involve taxonomies of NFQ categories, the information on particular details of those taxonomies is rather scattered. In our preliminary user study, described in Section 5.2.2, we tried to adopt an existing theoretical taxonomy [182] for complex questions, but the agreement on question categories between study coordinators was extremely poor, and did not improve even after a few rounds of discussion. For example, there is only a nuanced difference between the Causal Antecedent and Causal Consequent categories of that taxonomy. Thus, there was a need to gather information on existing NFQ taxonomies from all available sources and to create a taxonomy that is built with a transparent methodology and is thoroughly evaluated.

In this chapter, we aim to advance research in non-factoid question answering by examining the existing types of non-factoid questions, assessing their distribution in current datasets, and exploring the required answer forms.

We address several key research questions to achieve a comprehensive understanding of NFQA:

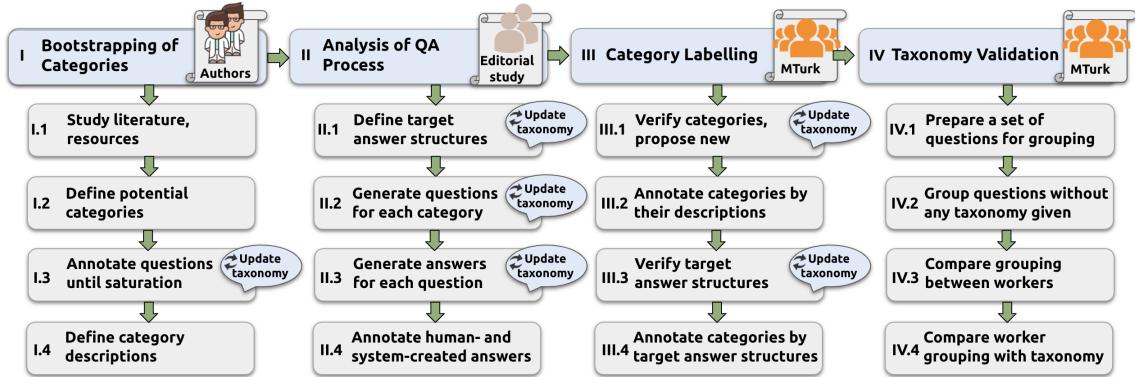
- Can we develop a new streamlined taxonomy of non-factoid questions that is both comprehensive and intuitive? How should it be validated to ensure that the taxonomy is neither as overly complex as existing theoretical models nor too simplistic, yet extensive enough to cover a broad range of non-factoid questions?
- Are there any categories that are particularly challenging to answer from a human perspective? How do system-generated and human-generated answers vary across different categories?
- Can we compile a new dataset of NFQ categories and establish a robust model for automatic category classification?
- What are the distributions of NFQ categories across various public QA datasets, and how does a state-of-the-art NFQA model perform across these categories?

5.2 Non-Factoid QA Taxonomy

The proposed taxonomy of NFQ categories and target answer structures was created through an iterative process shown in Figure 5.1. The final taxonomy, with examples, is given in Table 5.1. In this section, we describe each step of taxonomy creation and verification, and explore the most difficult-to-answer categories.

Table 5.1: The proposed taxonomy of NFQ categories and target answer structures.

Category	Description	Expected Answer Structure	Patterns
INSTRUCTION	You want to understand the procedure/method of doing/achieving something.	Instructions/guidelines provided in a step-by-step manner.	How to ...? How can I do ...? What is the process for ...? What is the best way to ...?
REASON	You want to find out reasons of/for something.	A list of reasons with evidence.	Why does ...? What is the reason for ...? What causes ...? How come ... happened?
EVIDENCE-BASED	You want to learn about the features/description/definition of a concept/idea/object/event.	Wikipedia-like passage describing/defining an event/object or its properties based only on facts.	What is ...? How does/do ... work? What are the properties of ...? What is the meaning of ...? How do you describe ...?
COMPARISON	You want to compare/contrast two or more things, understand their differences/similarities.	A list of key differences and/or similarities of something compared to another thing.	How is X ... to/from Y? What are the ... of X over Y? How does X ... against Y?
EXPERIENCE	You want to get advice or recommendations on a particular topic.	Advantages, disadvantages, and main features of an entity (product, event, person, etc) summarised from personal experiences.	Would you recommend ...? How do you like ...? What do you think about ...? Should I ...?
DEBATE	You want to debate on a hypothetical question (is someone right or wrong, is some event perceived positively or negatively?).	Arguments on a debatable topic consisting of different opinions on something supported or weakened by pros and cons of the topic in the question.	Does ... exist? Can ... be successful? Do you think ... are ...? Is ... really a ...?

Figure 5.1: Taxonomy creation procedure.

5.2.1 Bootstrapping of Categories

To create a draft of the NFQ taxonomy, an initial set of categories came from studying the literature on NFQA and researching web-resources dedicated to specific categories of NFQs such as [wikihow.com](https://www.wikihow.com/)¹, [debate.org](https://www.debate.org/)², and [differen.com](https://www.differen.com/)³. First, we identified a set of disjoint question categories covering most of the questions. Then, a set of questions was assessed to refine the categories and evaluate the comprehensiveness of the preliminary taxonomy. Questions were randomly sampled from each of the datasets NFL6 [43], MS MARCO [140], ELI5 [59], PhotoshopQA [54], SubjQA [17], StackExchange⁴, and Quora Question Pairs⁵ and crawled dataset from Kialo⁶. As we focused on self-contained non-factoid questions (such as web-search queries), we did not include the NLQuAD dataset where most questions depend on external context in order to be understood and disambiguated [173]. In total, 800 questions were assessed before reaching saturation, as no new categories or ambiguities emerged further in the assessment.

5.2.2 Analysis of QA Process

Following the development of a draft taxonomy of non-factoid question categories, the subsequent step involved determining the target answer structures and understanding the complexity of answer generation for each category. This analysis was conducted through a controlled editorial user study, as described in the previous Chapter (see Section 4.3). This section will focus on aspects of the study process and results that relate to the analysis of question and answer generation for each category.

¹<https://www.wikihow.com/>

²<https://www.debate.org/>

³<https://www.differen.com/>

⁴<https://archive.org/details/stackexchange>

⁵<https://www.kaggle.com/c/quora-question-pairs/data>

⁶<https://www.kialo.com>

Table 5.2: Examples of evaluated questions and human/system answers from the editorial user study.

Category	Generated question	System answer	Human answer
INSTR.	How can I change the link colour in Elementor? <i>(Easy to answer: 4.0)</i>	Select a section and then switch to style. Now you see an option typography. Just click to open that. And now you notice you have an option link color right here. <i>(Perceived usefulness: 3.0)</i>	Click on the 3 lines on the top left-hand corner, click "Theme Style" and then click "Typography". Then go to the "Link" heading and from there you can change its colour! <i>(Perceived usefulness: 3.0)</i>
REASON	Why are my Peace Lily flowers turning brown? <i>(Easy to answer: 4.0)</i>	Peace lily blossoms change color as they age. ... After about 10 days, peace lily blossoms turn green again. Pollination and seed setting also cause the flowers to turn green. The flowers usually stay on the plant for about a month before they begin to die and turn brown, and finally they wither and fade. <i>(Perceived usefulness: 4.0)</i>	All flowers die but you are probably may be over or under watering. Most likely overwatering. <i>(Perceived usefulness: 3.0)</i>
EVID.	What are the benefits of topically applying Vitamin E? <i>(Easy to answer: 4.0)</i>	Vitamin E is a powerful antioxidant that may be effective at reducing UV damage in skin. And vitamin E applied topically may help nourish and protect your skin from damage caused by free radicals. <i>(Perceived usefulness: 4.0)</i>	It can repair skin and lighten dark spots but you can also have too much of it. <i>(Perceived usefulness: 1.0)</i>
COMP.	How are organic vegetable seeds better than regular seeds? <i>(Easy to answer: 4.0)</i>	Organic Seeds Are More Robust Seeds and plants do not know they are being grown organically. ... The genetics of the seed does not change after growing plants organically for a few years. Good quality seeds from organic farms or conventional farms will grow equally well in your soil. <i>(Perceived usefulness: 4.0)</i>	There really is no difference between them! It's mostly marketing spin. <i>(Perceived usefulness: 0.0)</i>
EXP.	Where is the best Italian restaurant in [LOC] ⁷ ? <i>(Easy to answer: 2.0)</i>	Best Italian Restaurant in [LOC] - Menu, Photos, Ratings and Reviews of Restaurants serving Best Italian in [LOC]. Best [LOC] Italian. <i>(Perceived usefulness: 0.0)</i>	Da Noi is rated 4/6 on Zomato website. <i>(Perceived usefulness: 4.0)</i>
DEBATE	Does god exist? <i>(Easy to answer: 2.0)</i>	There remain many mysteries that are beyond science. Does that mean that a God truly exists? A scholar gives reasons for this possibility. <i>(Perceived usefulness: 0.0)</i>	Many believe he does and many believe he doesn't. It's up to you to make up your mind about whether you believe God exists. <i>(Perceived usefulness: 1.0)</i>

During discussions with the study participants, we further refined the taxonomy by establishing target answer structures and reformulating category descriptions based on clarification questions asked by the participants. Some categories were renamed to better reflect the target answer structures. The reasons were twofold: firstly, to align our taxonomy more closely with factoid question taxonomies, which typically name categories after the type of entity expected in the answer — PERSON, LOCATION, NUMBER, etc.; and secondly, to prevent any bias that might lead annotators to expect questions of certain categories to start with specific words — thus, we changed HOW-TO to INSTRUCTION and WHY to REASON. The preliminary expected answer structures for each category were developed during the trial run of the study, as outlined in the study procedure section (see Section 4.3.1).

Table 5.2 showcases examples of questions and answers for each category, generated during our study. It includes questions crafted by askers with their perceived difficulty ratings from the first stage of the study, answers generated by other participants (answerers) with predicted usefulness from the answerers' perspective in the second stage, and usefulness assessments from the askers in the final, third stage of the study.

Table 5.3 provides a breakdown of the perceived difficulty of various non-factoid question categories and the corresponding usefulness of human and system-generated answers. Notably, the system performance varies greatly between the categories.

The results highlight that FACTOID and EVIDENCE-BASED questions, which generally require specific factual answers, were rated as the easiest to answer. In contrast, categories such as DEBATE and EXPERIENCE, which often require nuanced understanding or personal insights, are marked as more challenging.

System-generated answers, mainly from Google SERP snippets, excelled in categories where definitive, concrete answers are expected. However, in more subjective or intricate categories like DEBATE, EXPERIENCE and REASON, human-generated answers were perceived as more useful. For the DEBATE category, the difference between system- and human-generated answer usefulness was statistically significant (Student's t-test, $p < 0.05$). This highlights a gap in the system's ability to handle questions requiring deep understanding and personalization — which remains a significant challenge in the field of non-factoid question answering.

Table 5.3: Perceived difficulty of questions on scale from 0 (very likely to answer) to 2 (not likely to answer); and the usefulness of corresponding human and system answers, rated from 0 (not useful) to 4 (very useful).

Category	Question difficulty	Answer usefulness	
		system	human
INSTRUCTION	0.27	2.44	2.40
REASON	0.54	2.15	2.69
EVIDENCE-BASED	0.15	2.71	2.29
COMPARISON	0.29	2.40	2.54
EXPERIENCE	0.50	1.96	2.42
DEBATE*	1.00	1.62	2.29
FACTOID	0.02	3.02	2.69

* significant difference between system/human answers.

5.2.3 Category Labelling

To ensure the reliability and applicability of our newly developed taxonomy of non-factoid questions, we proceeded to validate the category labeling process and curate a comprehensive dataset. This was achieved through a series of larger-scale crowdsourcing editorial studies conducted on Amazon Mechanical Turk (MTurk). For all crowdsourcing tasks (HITs), workers were selected following best practices for data collection on MTurk: HIT approval >95%, HITs approved >100 [148]. After several trial tasks for each study, where workers were able to provide feedback regarding the clarity of the task or any concerns about the reward amount, the best-performing workers were selected and assigned a special

qualification type for them to continue the actual editorial study, to ensure that it was completed by more reliable assessors. Workers were allowed to participate in only one type of a crowdsourced study, to exclude possible bias from already being familiar with the taxonomy. Each HIT was assessed by 3 workers in all experiments. The crowdsourcing studies were reviewed and approved by the Human Research Ethics Committee of the RMIT University⁸.

Question Category Labelling

First, to verify how well people understand the question categories based on their descriptions, and whether they have issues with choosing a category, we designed a crowdsourcing study for non-factoid question category labelling. In each HIT, workers were shown three questions and seven categories: six from the proposed taxonomy plus one additional OTHER/MULTI category for questions that do not fall into any other category, or fall into multiple categories. The categories appeared under each question in a random order, to prevent potential positioning bias. Workers were instructed to read each category description given in the instructions, and choose the most appropriate category for each question. For convenience, a shortened version of the category description appeared when the respective category name was clicked on. For the OTHER/MULTI category, workers were given the option to provide their own category name/description. Until pressing the submit button, workers were free to change the chosen category for each question. During trial annotation runs, all cases where a question received three different labels or suggestions for OTHER/MULTI category were reviewed, and the necessary changes were made to improve the taxonomy and category descriptions.

One out of three questions in each task was a gold question with a known answer, previously assessed during the taxonomy bootstrapping stage 5.2.1 and also given the same label by ≥ 2 workers during trial runs. In total, we had 273 gold questions. A HIT was automatically approved when the gold question was answered correctly, and rejected otherwise. We manually studied all automatic rejections and refunded them if workers had good justifications for their annotation in the rejection form. Each HIT was rewarded with \$0.2 upon approval. At the end of this stage, some question category names and descriptions were simplified and clarified based on worker comments.

In total, 1000 questions were assessed by at least three workers in this study, and the final question category was chosen by majority voting. The inter-annotator agreement between assessors was moderate with 0.54 Fleiss' kappa [63], which should also be interpreted in the context of the relatively high number of categories.

⁸The study was reviewed and approved by the Human Research Ethics Committee of the RMIT University. The IRB approval number is 22334.

To additionally evaluate the quality of assessment and comprehensibility of categories, pairs of duplicate questions from the Quora Question Pairs dataset⁹ were mixed into different HITs, based on the assumption that two questions which simply paraphrase each other should belong to the same category. Only 4 pairs out of 154 received inconsistent labels, demonstrating that workers generally agreed on categories for paraphrased questions.

Target Answer Structure Evaluation

To verify how well the expected answer structure describes its corresponding category, we ran another round of crowdsourcing assessments using the same interface, except showing target answer structure descriptions instead of category descriptions. After completing a number of trial runs and selecting assessors, we refined the target answer structures based on workers' comments on task clarity. We used the same approval process as in the previous study, based on gold questions and subsequent manual inspection of rejections. At least three workers assessed 850 questions, with inter-annotator agreement being moderate (0.53 Fleiss' kappa), showing that both category and answer descriptions are equally understandable and suitable for assessment.

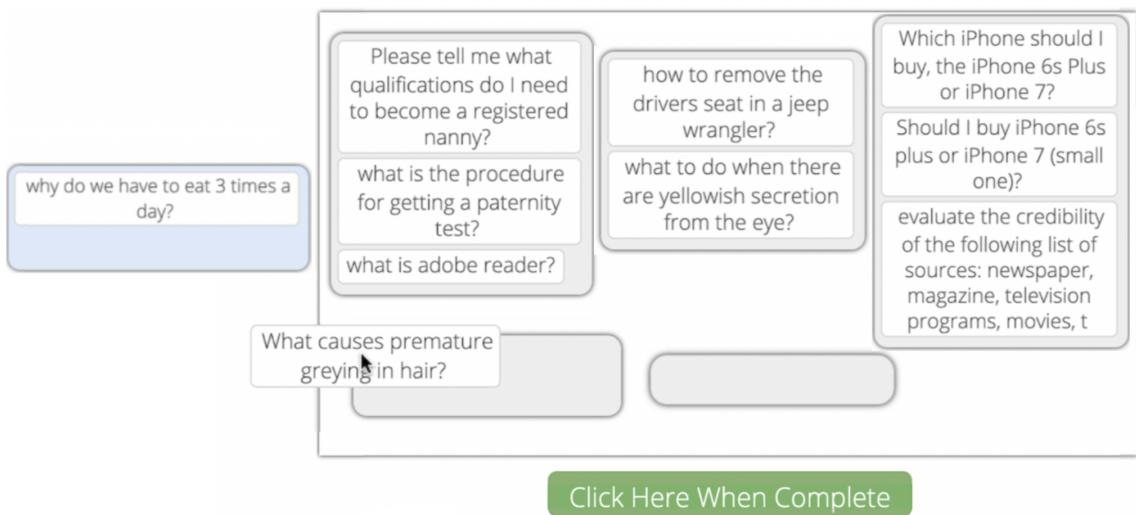
In total, for both studies, 12.5% of HITs were automatically rejected, of which 5% had good justifications and were ultimately refunded. Only 7 questions were labelled as the MULTI category, falling under two or more categories; 8 questions received 3 different labels 7 of which were assigned MULTI label. The final taxonomy category names, descriptions, and target answer structures are given in Table 5.1.

5.2.4 Taxonomy Validation

To study how well our taxonomy corresponds to how people naturally group questions, we conducted an exploratory crowdsourcing study. In this study, crowdsource workers were asked to group a set of questions by type or the expected form of an answer. Importantly, these participants were not given any guidance (e.g. a taxonomy, or examples) and were free to define the groups as they saw fit. The procedure for choosing reliable assessors, and clarification of the task, was the same as described in Section 5.2.3. The questions that participants were asked to group came from the set of questions that were assessed during the previous editorial and crowdsourcing studies.

Here, we aimed to imitate the process of category bootstrapping (Section 5.2.1) on a bigger scale and compare the results to the new taxonomy. We designed an interface, shown in Figure 5.2, where questions from the blue area on the left were dragged and dropped into the white box on the right to form a new grey box or expand one of existing grey boxes

⁹<https://www.kaggle.com/c/quora-question-pairs/data>

Figure 5.2: Interface for the question clustering study.

(question groups). Before clicking the submit button, workers could change their grouping arbitrarily. At the beginning of the task, workers saw ten questions to be arranged and no pre-defined gray boxes. Since it might have felt more natural for participants to attempt to group questions based on topics, rather than on question forms as required, we constructed each set of questions in the task so as to include as many different topics as possible, by clustering the whole set of questions in the dataset prior to the study.

We chose HDBSCAN¹⁰ [30] as the clustering algorithm due to its robustness to noise. The questions were featurised using Universal Sentence Encoder (USE)¹¹ [117]. The parameters “min_cluster_size” and “min_samples” were set to 2 for HDBSCAN, with the default values used for other parameters. We obtained 1261 clusters in total.

For each HIT we randomly sampled eight questions from different clusters, reducing the likelihood of paraphrased questions or similar topics occurring in a set. The remaining two (out of ten) questions in each HIT were previously annotated paraphrased questions from the Quora Question Pairs dataset. A HIT was automatically approved if the following conditions were met:

1. the number of submitted question groups was > 1 ;
2. the two paraphrased questions from Quora were allocated into the same group.

The first check was enabled by ensuring that each question set contains more than one topic and more than one category. The second check was supported by the idea that two paraphrased questions with the same category label should fall into the same group

¹⁰https://hdbSCAN.readthedocs.io/en/latest/how_hdbSCAN_works.html

¹¹<https://tfhub.dev/google/universal-sentence-encoder-multilingual/3>

regardless of the grouping logic. Each of the 36 HITs (360 questions/72 gold items) was completed by three workers; only non-gold questions are used in all subsequent analysis. The workers were rewarded \$0.5 for each approved HIT.

Inter-Participant Cluster Similarity

To investigate the level of similarity between the questions groups (further referred to as clusters) created by different participants, we study the agreement between workers, framing the clustering problem as a binary classification task [9]. Each pair of questions receives label 1 if these questions were assigned to the same group, and label 0 otherwise. After this transformation, we can calculate the agreement between participants using Fleiss' kappa; the workers had almost no agreement with a kappa value of 0.05. Given the sophisticated HIT approval process, we attribute the absence of agreement to the task of unsupervised clustering being much more challenging than labelling in accordance with an existing categorization. Manual verification of 150 random groups confirmed that the workers did create logical groupings of questions.

Clustering and Taxonomy Similarity

To evaluate the similarity between the natural groupings in this task and taxonomy categories, for each question set in a HIT, clusters created by workers in the current experiment can be compared to the “reference” clusters created by aggregating questions using category labels previously assessed by the question classification crowdsourcing study. This is quantified using the V-measure [159], which compares clusters with reference clustering in terms of homogeneity and completeness. A perfectly homogeneous clustering is one where each cluster has data points belonging to the same class label. In our case, homogeneity measures how uniformly each cluster contains only questions from a single category. A perfectly complete clustering is one where all data points belonging to the same class are clustered into the same cluster. In our case, completeness assesses how well all questions from the same category are assigned to the same cluster. The V-measure score was 0.6 (with homogeneity of 0.73 and completeness of 0.55). Reference clustering based on the taxonomy on average consisted of 3.6 clusters (i.e. on average each set of 8 questions contained 3.6 non-factoid categories) while workers recorded a mean of 5.5 clusters. The relatively high homogeneity score shows that workers usually grouped questions together similarly to our taxonomy categorization, but their clustering was slightly more fine-grained.

The results of the study show that people with no prior knowledge of our question taxonomy naturally tend to place questions of the same taxonomy category together in one group; however, their groups are typically smaller, and vary substantially from person to

person. Manual data inspection showed that assessors in our study seemed to be using some additional individual rules for creating sub-categories. A random example of grouping performed by MTurk workers is presented in Table 5.4, with the reference clustering based on the taxonomy displayed at the bottom.

Throughout this section, we have described the long process of careful taxonomy construction and its detailed verification. First, after bootstrapping, the taxonomy was refined in an editorial user study where target answer structures for each category were created and the most challenging categories were identified. According to the findings, people consider DEBATE, REASON and EXPERIENCE questions to be the hardest to answer. System-generated answers for the same categories have the lowest answer usefulness score, falling far behind human-generated answers. A study using question clustering showed that the proposed categorization reflects how people naturally group questions without any guidance. This, alongside moderate agreement between assessors in question categorization in accordance with the taxonomy, gives us confidence that the taxonomy categories are well-defined.

5.3 Question Category Prediction

The labeled data obtained from the bootstrapping stage and collected during MTurk studies have been compiled into a dataset called NF-CATS. This section describes this new dataset and how we utilized it for the task of question category prediction, conducting experiments with different classification models.

5.3.1 NF-CATS Dataset

The NF-CATS dataset contains examples of natural questions divided into categories from our taxonomy and two supplementary categories: FACTOID (questions that require a short factual answer) and NOT-A-QUESTION (sentences without question intent). The supplementary categories are introduced in order to facilitate the training of models that can predict question categories based on any given text. Questions from these categories were mostly collected via unsupervised means, the details of which are provided below. A large portion of diverse DEBATE questions was also obtained in an unsupervised fashion from Kialo¹² (a web-resource that specialises in debates). For each question in our dataset, we provide a column that indicates the source of the assessment: (1) *MTurk*: the category was annotated by three MTurk workers and they reached an agreement; (2) *Bootstrapping*: the category was annotated during the bootstrapping stage; (3) *Auto*: the category was

¹²<https://www.kialo.com>

Table 5.4: Random examples of questions grouped by workers.

	Clusters
worker #1	<p>what is a cultivator?</p> <p>what is adrenogenital syndrome?</p> <p>how do i burn dvd's using window's media player?</p> <p>how can i show messege box in web based c#?</p> <p>How do I offset irregular periods and get back on a regular menstrual cycle?</p> <p>How expensive is it to call finland?</p> <p>What happened to the Greek Gods and Goddess?</p> <p>how do you get illeagls to get out of the u.s.?</p>
worker #2	<p>what is a cultivator?</p> <p>What happened to the Greek Gods and Goddess?</p> <p>how do i burn dvd's using window's media player?</p> <p>how can i show messege box in web based c#?</p> <p>how expensive is it to call finland?</p> <p>how do you get illeagls to get out of the u.s.?</p> <p>What is adrenogenital syndrome?</p> <p>How do I offset irregular periods and get back on a regular menstrual cycle?</p>
worker #3	<p>what is a cultivator?</p> <p>how do i burn dvd's using window's media player?</p> <p>how do you get illeagls to get out of the u.s.?</p> <p>How do I offset irregular periods and get back on a regular menstrual cycle?</p> <p>How expensive is it to call finland?</p> <p>How can i show messege box in web based c#?</p> <p>What happened to the Greek Gods and Goddess?</p> <p>what is adrenogenital syndrome?</p>
proposed NFQA taxonomy	<p>what is a cultivator?</p> <p>What happened to the Greek Gods and Goddess?</p> <p>how expensive is it to call finland?</p> <p>what is adrenogenital syndrome?</p> <p>how do you get illeagls to get out of the u.s.?</p> <p>How do i burn dvd's using window's media player?</p> <p>How can i show messege box in web based c#?</p> <p>How do I offset irregular periods and get back on a regular menstrual cycle?</p>

Table 5.5: Breakdown of NF-CATS dataset.

Category	Bootstrap.	MTurk	Auto	Total
INSTRUCTION	132	413		545
REASON	119	166		285
EVIDENCE-BASED	325	863		1188
COMPARISON	21	62		83
EXPERIENCE	70	75		145
DEBATE	93	99	1224	1416
FACTOID	34		3822	3856
NOT-A-QUESTION			4466	4466
Total	794	1678	9512	11984

assigned in an unsupervised fashion based on the question source (e.g. a web-resource that specializes in a particular category).

The breakdown of the categories and the assessment sources is shown in Table 5.5, and train/validation/test splits of the dataset are presented in Table 5.6. The dataset is imbalanced, with some categories rarely appearing in the annotated part and others being substantially augmented with unsupervised data. Suitable unsupervised sources of questions for under-represented categories remain to be found. We do not expect the exploratory analysis in Section 5.4 to be significantly affected by the class imbalance.

Supplementary Categories

When expanding our dataset, it's crucial to include classes that represent "unknown" or "other" examples to help the model generalize better across a wider variety of inputs. In NF-CATS, we incorporated two supplementary categories: FACTOID and NOT-A-QUESTION. FACTOID includes questions that can be answered with short, factual responses, while NOT-A-QUESTION encompasses sentences that lack a questioning intent.

For the FACTOID category, we sourced questions from three QA datasets that mostly contain factoid questions: TweetQA [204], BoolQ [38], and the development split of SQuAD [156]. To ensure that NFQs were not included, we only extracted questions from SQuAD and TweetQA that satisfy two requirements:

1. The answer contains fewer than four words.
2. Spacy NER¹³ detects at least one named entity or a numerical/year entity in the answer.

Given the simplicity of “Yes/No” questions contained in the BoolQ dataset, we randomly sampled 1,500 questions from it with no further filtering.

¹³<https://spacy.io/usage/linguistic-features#named-entities>

To populate the NOT-A-QUESTION category, sentences were extracted from document contexts of SQuAD and TweetQA questions which had been selected as FACTOID in the previous step. Each context was split into sentences, and all sentences ending with a question mark were excluded. This approach ensures that the dataset includes declarative sentences closely related to the factual questions, increasing the difficulty for the models in differentiating between these categories.

5.3.2 Question Category Classification

To explore the performance of different approaches in the task of question category prediction, we use four models. Per-category and macro F1-scores for the models are shown in Table 5.6.

First, we use logistic regression over *tf-idf* feature vectors as the baseline classification model. The baseline model was implemented in the Scikit-Learn framework¹⁴. To automate the selection of high-impact hyper-parameters such as regularization strength and vocabulary size, we use the Optuna [4] hyper-tuning framework with macro F1-score on the validation set as the objective. To find the best hyper-parameters, 1000 trials of hyper-tuning search were executed.

We also leveraged three Transformer models: BERT-base [51], RoBERTa-base [127], and RoBERTa-base fine-tuned on SQuAD2.0¹⁵. The latter was chosen due to its potentially better domain fit for this task. Each Transformer network was followed by 2 feed-forward layers with Mish activation [134] and a classifier layer on top. Cross-entropy was used as the loss function, and AdamW as the optimizer. As our dataset is imbalanced, we applied batch balancing when training Transformer models, sampling N=8 random examples of each class for each mini-batch. The models were implemented using the AllenNLP framework¹⁶ [67]. The training took up to 10 hours on a single NVidia Tesla P100 16GB GPU. Hyper-parameters were selected manually based on the validation loss, with 10 runs, and shared across all models. Weighted F1-scores on the validation set for the best epochs were equal to 0.954 (5th epoch), 0.957 (4th epoch), and 0.958 (6th epoch) for BERT-base, RoBERTa-base, and RoBERTa-SQuAD2.0, respectively.

As expected, Transformer models provide a substantial gain in performance over the simpler linear baseline model, with RoBERTa-SQuAD performing the best. Even though RoBERTa-SQuAD only moderately outperformed BERT-base and RoBERTa-base in terms of macro F1-scores, we found it to have much better generalisation and higher robustness when manually evaluating predictions on different NFQA datasets, as described in

¹⁴<https://scikit-learn.org/stable/>

¹⁵<https://huggingface.co/deepset/roberta-base-squad2>

¹⁶<https://allennlp.org/>

Table 5.6: Performance of classifier models on NF-CATS test set.

Category	LOGREG	BERT	RoBERTA	RoBERTA	DATASET SIZE		
	TF-IDF	BASE	BASE	SQuAD2.0	Test	Train	Val
INSTRUCTION	0.856	0.916	0.917	0.943*	113	346	86
REASON	0.849	0.893	0.852	0.893	59	181	45
EVIDENCE-BASED	0.860	0.906	0.918	0.946*	237	761	190
COMPARISON	0.750	0.741	0.815	0.828	15	54	14
EXPERIENCE	0.612	0.566	0.528	0.653	26	95	24
DEBATE	0.911	0.952	0.954	0.957	283	906	227
FACTOID	0.954	0.981	0.980	0.987*	771	2468	617
NOT-A-QUESTION	0.993	0.998	0.998	0.997	893	2858	715
Macro F1-score / Total	0.848	0.869	0.870	0.901*	2397	7669	1918

* significantly different from the TF-IDF baseline. (Student’s t-test, $p < 0.05$).

Section 5.4. For instance, BERT-base had a notable skew towards the INSTRUCTION category in its predictions, resulting in many visible false positives.

5.4 Benchmarking Categories on QA Datasets

In this section, we examine the distribution of questions across the categories defined in our taxonomy within various public QA datasets. This analysis will help identify potential biases toward certain categories and provide insight into the representation of various question types. Additionally, we evaluate the performance of a state-of-the-art NFQA system trained on one of these datasets. We aim to assess how well the system performs across different question categories and to compare the system’s per-category performance with the usefulness ratings of search engine result page (SERP) snippets, as outlined in Table 5.3. This analysis will allow us to understand any discrepancies between system performance and user-perceived answer usefulness, highlighting areas for potential improvement in NFQA systems.

5.4.1 Analysis of Category Distribution

Table 5.7 shows the distributions of categories in four datasets, based on predictions of the top performing RoBERTa-SQuAD model from Section 5.3.2. We evaluate the dataset for question classification compiled by Li and Roth. The original annotation features 78.2% FACTOID questions and 21.8% NFQs, with 13.8% roughly mapping to the EVIDENCE-BASED category from our taxonomy. Model predictions highlight the same predominance

Table 5.7: Question category distributions in datasets.

Categories	Datasets			
	Li&Roth TREC	MS MARCO Bing queries	NFL6 Yahoo answers	ELI5 (Test set) Reddit ELI5
INSTRUCTION	3.47%	6.00%	40.33%	9.80%
REASON	2.18%	2.31%	24.85%	45.80%
EVIDENCE-BASED	25.29%	34.52%	23.92%	24.20%
COMPARISON	0.65%	0.39%	3.42%	1.80%
EXPERIENCE	1.23%	0.21%	2.03%	0.60%
DEBATE	0.55%	0.96%	2.70%	8.20%
FACTOID	64.40%	55.21%	2.57%	5.50%
NOT-A-QUESTION	2.23%	0.40%	0.19%	4.10%
TOTAL	5,894	1,026,758	24,512	87,361

of categories that require factual answers, classifying 89.69% of the dataset into either FACTOID or EVIDENCE-BASED. The manual analysis uncovered a few false positives in rare categories, hinting at those categories having an even smaller representation in reality.

To investigate the distribution of questions submitted to web-search engines, we study the MS MARCO dataset [140], which contains more than 1,000,000 user search queries submitted to Microsoft Bing. According to our analysis, most of the queries are either FACTOID or EVIDENCE-BASED, with the INSTRUCTION category being represented to a smaller degree. This suggests that the performance of models displayed on the MS MARCO leaderboard¹⁷ is mainly a reflection of the ability of systems to answer FACTOID and EVIDENCE-BASED questions. It is arguable whether people rarely ask more sophisticated categories of questions due to a smaller need, or simply because they do not expect current systems to answer them. The existence of dedicated web-resources for DEBATE (debate.org¹⁸, kialo.com¹⁹), INSTRUCTION (wikihow.com²⁰), and COMPARISON (diffen.com²¹) might point towards the second explanation, necessitating further research of under-represented categories.

Finally, we analyse the open-domain NFQA datasets NFL6 and ELI5, both based on data from CQA platforms. In the NFL6 dataset derived from the Yahoo’s Webscope L6 collection²², the largest category is INSTRUCTION, followed by EVIDENCE-BASED and REASON. Together with the extremely small representation of INSTRUCTION questions in MS MARCO, this might suggest that people prefer to use CQA platforms over web-search engines for various “how-to” questions. However, this bias may also be influenced by the

¹⁷<https://microsoft.github.io/MSMARCO-Document-Ranking-Submissions/leaderboard/>

¹⁸<https://www.debate.org/>

¹⁹<https://www.kialo.com>

²⁰<https://www.wikihow.com/>

²¹<https://www.diffen.com/>

²²<https://webscope.sandbox.yahoo.com/catalog.php?datatype=l>

specific time period when the data was collected and the filtering criteria applied.

The more narrowly focused ELI5 [59] dataset consists of 270K threads from the “Explain Like I’m Five” subreddit. Most of the questions in the dataset are from the REASON and EVIDENCE-BASED categories, representing information requests that require explanations. This indicates that the KILT ELI5 leaderboard²³ primarily reflects the performance of systems on just these two question categories.

For categories like COMPARISON, DEBATE, and EXPERIENCE, representation in public benchmarks is notably sparse. Particularly, the DEBATE and EXPERIENCE categories, which require nuanced and subjective answers, are among the most challenging. This suggests that current datasets may not fully capture the diversity of real-world information needs, especially for questions that demand more complex and elaborate responses.

5.4.2 NFQA Model Performance Across Categories

In Section 5.2.2, Table 5.3, we evaluated how well NFQ categories are answered by Google web-search snippets, and identified the most challenging NFQ categories to be DEBATE, EXPERIENCE, and REASON. Here, we focus on evaluating the performance of a state-of-the-art model specifically trained to answer NFQs, to understand the influence of the category imbalance in NFQA datasets and to determine challenging categories for the model. For this, we utilize the ELI5 dataset for abstractive long-form QA. Unfortunately, the unsupervised evaluation methodology for long-form NFQA adopted by Fan et al. [59], namely ROUGE score variants, is not representative of the model performance, to the point where randomly selected answers produce higher scores than ground truth answers [106]. Thus, we leave large-scale unsupervised evaluation for future work, along with the research of more suitable metrics for NFQA. Instead, we carried out A/B human evaluation in a form of a small-scale editorial study across different NFQ categories in ELI5, following the human evaluation methodology of Krishna et al. [106].

Questions and gold answers for assessment were sourced from the corrected evaluation split of ELI5 provided by Krishna et al. Model answers were generated using the best system of Krishna et al. (with $p = 0.9$) consisting of a “contrastive REALM” dense retriever and a generator based on the Routing Transformer, the current state-of-the-art model for representing long-range dependencies in sequences via sparse attention and mini-batch k -means clustering [160].

Volunteers were asked to select the “better” answer for one question at a time, choosing between gold and system-generated answers presented in random order and without labelling the source. Unlike the original A/B testing setup used by Krishna et al., in “Tie”

²³<https://eval.ai/web/challenges/challenge-page/689/leaderboard/1908>

Table 5.8: Per-category performance of a state-of-the-art model:
A/B human evaluation on ELI5.

Categories	System Performance	Prefer Gold	Prefer System	Both Good	Both Bad
INSTRUCTION	30% (6/20)	50% (10)	15% (3)	15% (3)	20% (4)
REASON	40% (8/20)	40% (8)	20% (4)	20% (4)	20% (4)
EVIDENCE-BASED	45% (9/20)	30% (6)	15% (3)	30% (6)	25% (5)
COMPARISON	40% (8/20)	45% (9)	10% (2)	30% (6)	15% (3)
EXPERIENCE	20% (4/20)	60% (12)	10% (2)	10% (2)	20% (4)
DEBATE	5% (1/20)	75% (15)	0% (0)	5% (1)	20% (4)
TOTAL	30% (36/120)	50% (60)	12% (14)	18% (22)	20% (24)

situations when both answers were equal in their quality (either good or bad), volunteers were instructed to select “Both Good” or “Both Bad” options, respectively. This change allowed us to evaluate the overall percentage of good answers given by the system. In total, we had 5 volunteers and 120 questions, with 20 questions per each NFQ category. All volunteers were English-speaking and had at least a Master’s degree. This minimal-risk editorial study involved only the evaluation of publicly available data. All volunteers were anonymized, and no personal information was collected. Question categories were assigned through MTurk evaluation in the same manner as described in Section 5.2.3. The results are presented in the Table 5.8.

On average, the system answers were preferred only in 12% of cases, which is slightly less than 14% reported for the system by Krishna et al. We attribute this to the difference in distributions of categories between our evaluation and the original evaluation. Krishna et al. randomly sampled questions, and the majority belonged to REASON and EVIDENCE-BASED categories, while we sampled questions uniformly across categories.

The “System Performance” column gives the overall system performance, which measures the percentage of system answers that were either preferred over gold answers or considered equally good. Similarly to the performance of the production-grade system tested in our editorial study (3rd column in Table 5.3), the most challenging category for the system trained on ELI5 is DEBATE, where the system gave only one good answer out of 20, followed by EXPERIENCE with four good answers. These two categories are poorly represented in the training data. On the other hand, the performance of the ELI5-trained system for the REASON category is relatively good, which could be explained by a very high representation of this question category. This supports our hypothesis that system performance may be affected by unbalanced question categories in training data, especially for more challenging categories.

5.5 Chapter Summary

This chapter has significantly advanced our understanding of non-factoid question answering by developing a comprehensive and streamlined taxonomy of non-factoid question categories, along with defining the expected answer forms for each category. This taxonomy, built on a robust methodology, has been validated through both editorial and crowdsourcing studies, ensuring its relevance and applicability in real-world scenarios.

We compiled the labeled data into a new dataset of NFQ categories, accompanied by a high-performance classifier. This tool is designed to help researchers categorize questions in other datasets, facilitating broader application and study of NFQA.

Our analysis of four existing QA datasets commonly used in NFQA research highlights a skewed distribution across the newly defined non-factoid question categories. This skewed distribution illustrates the limitations of current datasets, which fail to adequately cover the diverse spectrum of non-factoid questions, particularly under-represented categories such as DEBATE and EXPERIENCE. These findings emphasize the need for new datasets that encompass a wider range of question categories and for NFQA models capable of addressing the complexities of more challenging questions. Moreover, our evaluation of the current state-of-the-art NFQA systems revealed uneven performance across different categories, suggesting the necessity for more balanced benchmarks or category-specific benchmarks that focus on tailored evaluation methods to further advance these systems.

Building on these insights, the following chapter will concentrate on the development of a new NFQA dataset. This dataset will specifically target the challenging multi-document NFQA subtask within the INSTRUCTION category, which is popular yet under-represented and exhibits significant potential for improvement in system performance. In addition, we will enhance the human evaluation framework for NFQA, aiming to establish a robust benchmark that effectively measures system performance and fosters the development of more sophisticated NFQA systems.

Chapter 6

WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering

In the previous chapter, we established a taxonomy for categorizing non-factoid questions and revealed substantial imbalances in how these categories are represented within existing datasets. We also demonstrated that performance disparities exist among state-of-the-art systems across different non-factoid question types. These findings underscore the need for datasets specifically designed to address the unique challenges posed by different non-factoid question types.

As the final phase of this research, this chapter introduces WikiHowQA, a comprehensive benchmark designed for multi-document non-factoid question answering. This challenging subtask of NFQA requires reasoning across several documents to construct an answer. Developed from the WikiHow platform¹, which specializes in "how-to" instructions, this benchmark specifically targets the INSTRUCTION category. It comprises 11,746 human-written answers and 74,527 supporting documents, presenting unique challenges for current models. We describe the methodology used to construct the WikiHowQA benchmark, explore the challenges associated with multi-document non-factoid question answering, and establish strong baseline performances. Recognizing the complexities of evaluating non-factoid answers, particularly the challenges evaluators face with unfamiliar

¹wikihow.com/wikiHow:About-wikiHow

topics, we propose a novel human evaluation framework that leverages highlighted relevant supporting passages to enhance assessment accuracy.

This chapter aims to advance non-factoid question answering research by providing a robust benchmark for testing and developing models capable of effectively tackling the complexities of "how-to" questions.

6.1 Introduction

Non-factoid questions that require detailed, passage-level answers—such as explanations or opinions—pose significant challenges for current question-answering systems. Despite the existence of datasets designed for non-factoid question answering such as NFL6, and ELI5 [43, 59], our previous research (see Section 5.4.1) highlighted a notable imbalance in the coverage of different types of questions. In practice, users querying these complex questions often encounter only a summary of the most relevant document on a search engine results page (SERP). Such summaries are typically inadequate for addressing queries that necessitate sophisticated reasoning or synthesizing information scattered across multiple documents.

For example, straightforward questions like “how to fix my computer issue” might have multiple solutions, requiring the user to manually search through multiple retrieved documents to find the one that applies to their situation. More complex inquiries, such as “how to feel calm and relaxed” might demand an aggregated summary of the most popular methods from a variety of sources. Moreover, there are also questions for which answers have yet to be written, requiring a QA system to treat relevant documents as initial sources of information and then to reason out an answer based on them. For example, the question “how to prepare to buy a house in [neighbourhood], [city]” would necessitate retrieving specific information about local real estate markets and general home-buying advice, then reasoning through this data to construct a comprehensive answer.

The challenge extends to modern LLM-based QA models that, while capable of generating coherent and fluent responses, often include factually incorrect information, which can mislead non-expert users. These non-experts, typically the primary users of QA systems or annotators assessing system output, may struggle to identify subtle inaccuracies.

For example, when asking ChatGPT² the question “how to clean a plastic water tank” (Figure 6.1), the system provides well-structured and generally sound step-by-step instructions, including adding a cleaning solution and scrubbing the tank. However, it incorrectly states that the sanitization solution needs to sit for only about 30 minutes. In contrast, WikiHow³, a reputable source for instructional content, provides a different answer for the same question (Figure 6.2). WikiHow specifies a 24-hour sanitization period, consistent with its cited sources. This distinction highlights the importance of grounding QA systems in accurate information and providing transparent source citations to avoid misinformation. Subtle factual inaccuracies, such as the discrepancy in sanitization duration, may not be immediately apparent unless the user has specific expertise in the topic.

²ChatGPT is an advanced conversational AI system built on top of a state-of-the-art LLM, developed by OpenAI. More information is available at <https://chat.openai.com>.

³<https://bit.ly/About-wikiHow>

How to Clean a Plastic Water Tank

Cleaning a plastic water tank is important to keep the water fresh and prevent bacterial growth. Here are some general steps you can follow:

Materials Needed:

- Non-abrasive cleaners (like a mild dish soap)
- Long-handled brush or scrubbing tool
- Clean, non-abrasive cloth or sponge
- Water hose

Steps:

- Empty the Tank:** Start by draining or pumping all water out of the tank.
- Prepare a Cleaning Solution:** Mix a mild, non-abrasive cleaner with warm water. For
- Scrub the Inside:** Use a long-handled brush or scrubbing tool to clean the inside of the tank. Pay attention to corners and crevices where bacteria may accumulate.
- Rinse Thoroughly:** Once the tank is clean, flush it with clean water until there is no residual cleaner left.
- Sanitize the Tank:** If the tank holds drinking water, you may need to sanitize it after cleaning. Use a solution of household bleach and water (typically 1 tablespoon of bleach per gallon of water, but check with the tank manufacturer to be sure). Fill the tank with this solution, let it sit for about 30 minutes, then drain and rinse the tank thoroughly.
- Refill the Tank:** Once the tank is dry and clean, it can be refilled with water.

Always refer to your tank's specific care and maintenance instructions, as some materials or cleaning solutions may not be suitable for all tanks. Always use cleaning products that are

Figure 6.1: ChatGPT's response to the question "How to clean a plastic water tank." The instructions are well-structured but contain a factual inaccuracy regarding the recommended sanitization duration.

WATER > WATER CONSERVATION

How to Clean a Plastic Water Tank

Co-authored by wikiHow Staff and 16 contributors Last Updated: April 25, 2023 Approved

Article Summary

To clean a plastic water tank, open the outlet valve or tap and let the water flow out. Next, mix hot water with laundry detergent and use a bristle brush to scrub the inside of the tank with the cleaning solution to remove slime and sludge. Then, rinse out the tank thoroughly and flush the hoses and pipes. Finally, fill your tank with a mixture of clean water and bleach and let it sit for 24 hours to disinfect it. For tips on using a pressure washer to clean out your tank, read on!

Did this summary help you? Yes No

Help Categorize Articles! Help wikiHow readers by using this tool to confirm if our articles are in the correct categories.

Co-authored by: Madalaine Mason Apr 8, 2018 "We have just bought a second motor home."

corresponding source document:

Maine Center for Disease Control and Prevention
An Office of the Department of Health and Human Services

Cleaning, Disinfecting, a Flushing Drinking Water Storage Tanks at Small Public Water Systems

and/or sediments in the tank. Use a portable pump or drain to flush out dirty water from

STEP 2: Disinfecting
Use the disinfection equation to determine the amount of chlorine bleach to add to the tank that will result in a tank full of water with a free chlorine residual of 10 ppm. Add the calculated amount of bleach to the empty tank and fill the tank to the overflow level with water. Leave the tank filled for 24 hours.

Disinfection equation for a 10ppm free chlorine
→ Using 8.5% sodium hypochlorite
$$\text{Tank volume (gallons)} \times 10\text{ppm}$$
$$1,000,000 \times 0.085$$

Example for a 5,000 gallon tank

Figure 6.2: WikiHow's response to the question "How to clean a plastic water tank.", specifying a correct 20-hour sanitization period and citing reputable sources.

To address these challenges and move towards more advanced QA systems that can provide in-depth and factually accurate answers to a wide range of questions, we propose a new benchmark for the task of multi-document non-factoid QA (MD-NFQA). This task involves using multiple relevant documents to generate complete and factually grounded answers to non-factoid questions. Unlike many existing open-domain QA benchmarks that are biased toward passage-level reasoning and fail to account for document-level cues [97], the new benchmark is explicitly designed to evaluate models’ ability to synthesize information across multiple documents to better reflect the complexities of real-world QA scenarios.

We focus our benchmark on the INSTRUCTION category of non-factoid questions, which often begin with “how to”. These questions are underrepresented in current multi-document QA [183] datasets, despite their popularity, as evidenced by a 140% increase in “how to...” searches on Google since 2004⁴ and making up over half of the most searched queries.⁵

Our benchmark leverages WikiHow, a web-resource for INSTRUCTION questions containing over 235,000 articles on a wide range of topics. These articles provide comprehensive step-by-step instructions and are written by a community of experts and reviewed by an average of 16 people. The WikiHow website has proven to be a valuable resource for machine learning tasks [7, 16, 20, 45, 105, 113, 210, 214, 216], and in our work, we leverage WikiHow to create a high-quality benchmark specifically designed for MD-NFQA within the INSTRUCTION question category. Our benchmark consists of 11,746 questions from the INSTRUCTION category, each paired with a corresponding human-written answer, sourced from a diverse range of WikiHow articles. Each QA pair is supported by corresponding parsed relevant HTML pages from which the answer can be derived (Fig. 6.3). We evaluate several baseline models on the new benchmark that could serve as lower and upper bounds for model performance.

Human annotation is often used as the standard for evaluating long-form answers, but research has shown that individuals without specific training can only distinguish between human-generated and auto-generated text at a level equivalent to random chance [39]. This presents a challenge for evaluating the new benchmark for NFQA, particularly when the answers are lengthy or the topic is unfamiliar [106] or not interesting (see Section 3.3.1). Given that standard metrics struggle to detect factual inconsistencies, such as number swapping, negation, etc. [107], we delve into human sensitivity to such discrepancies by conducting a series of crowdsourcing experiments. Inspired by the elaborate yet resource-intensive manual evaluation framework of abstractive models proposed by Dou et al. [53],

⁴<https://bit.ly/telegraph-google-how-to>

⁵<https://bit.ly/most-asked-questions-on-google>

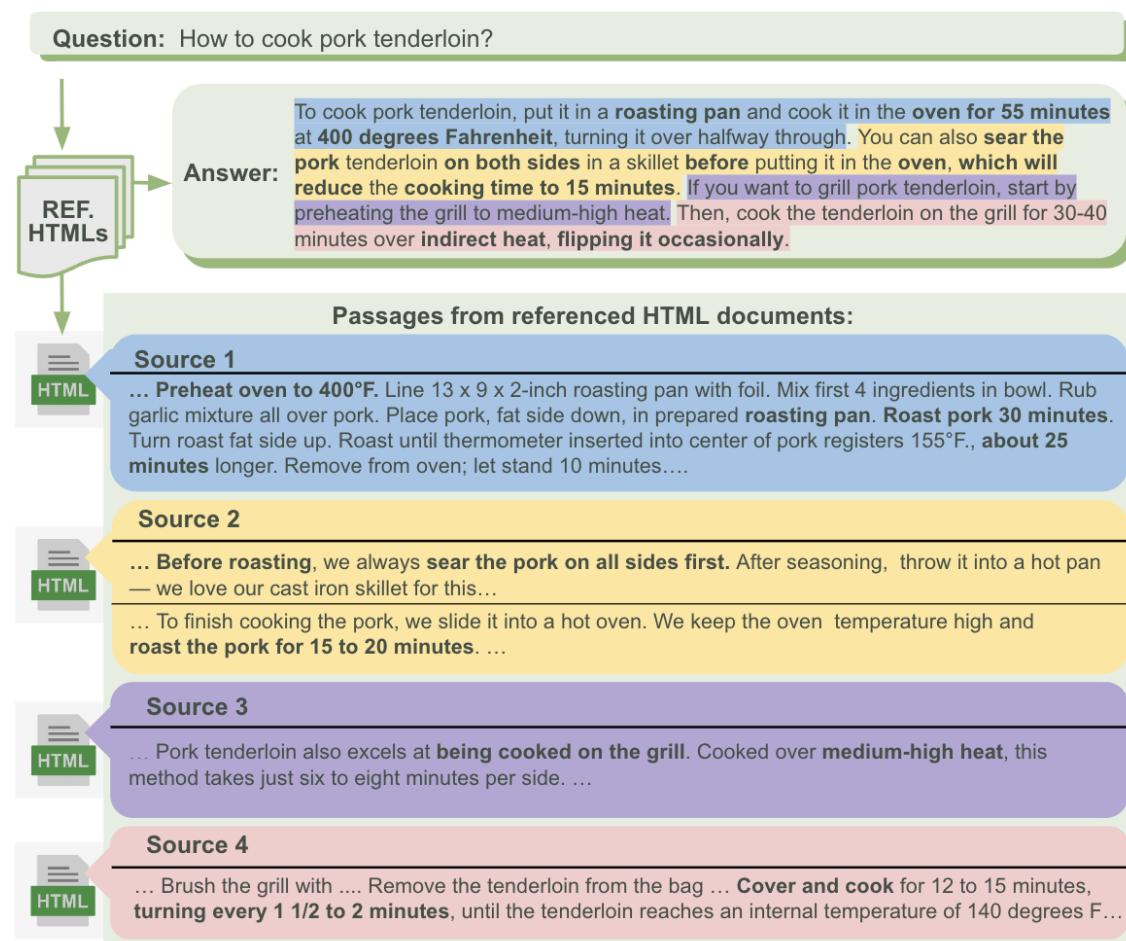


Figure 6.3: An instance of developed WIKIHowQA.

we introduce a simple human evaluation framework that leverages highlighted relevant passages to enhance the quality of NFQA assessments. Our findings demonstrate that incorporating highlighted relevant supporting passages into the evaluation process aids evaluators in understanding the context of each question and answer and improves the factual accuracy of the evaluation. As a result, we integrate this method into the evaluation of the newly developed benchmark.

The new benchmark for MD-NFQA serves as a valuable resource for the development of more advanced QA systems that can provide in-depth and comprehensive answers grounding information in supporting documents. By including parsed passages from relevant HTML pages and providing human judgments, our benchmark also has the potential to facilitate research in evaluating the factual correctness of long-form answers. All data is publicly available⁶ on the dataset website.⁷

⁶Under RMIT University DTA license for research use

⁷<https://lurunchik.github.io/WikiHowNFQA/>

RODENTS » CHINCHILLAS

How to Train a Chinchilla Question

Co-authored by Pippa Elliott, MRCVS 

Last Updated: March 29, 2019

Chinchillas are small rodents that are native to South America. They tend to be very energetic animals...

Manually written answer-summary

Method 1 Training Your Pet to Come to You

- 1** Choose the right treats. Treats are what will make or break this training exercise. You'll need to find ...
- 2** Present a treat. Your chinchilla ... training, which may take time and patience.^[1]
 - Hold the treat in one hand and lay your other open hand with your palm facing ...
 - Your open hand should act as a stepping stone for your chinchilla to climb onto ...
- 3** Try it without treats. Once your chinchilla learns ... time and practice.^[1]

Method 2 Teaching Your Chinchilla to Use a Litter Box

- 1** Note where your chinchilla urinates. Most chinchillas ... urinates.^[2]
- 2** Place the litter box where your pet urinates. If your ...

References 1. ↑ http://www.2ndchance.info/chinchilla.htm 2. ↑ http://www.pethealthandcare.com/chinchillas/chinchillas-training.html 3. ...	List of all references to source documents created by an author
---	--

Figure 6.4: An example WikiHow article.

6.2 Resource Description

This section presents our new resource, including its construction, text statistics, comparison with other benchmarks, and thorough quality analysis.

6.2.1 Data Collection

Fig. 6.4 illustrates an example of a WikiHow article accompanied by high-quality human-written article summary, which are equivalent to shorter passage-level answers to a question. To build WikiHowQA, we first downloaded over 236,000 articles published on WikiHow before January 2022 using the MediaWiki API⁸. We then filtered out articles without references, resulting in a collection of 74,527 articles. Among these articles, only about 20% had human-written article summaries, which we used as target answers. We downloaded the HTML content of all cited URLs using the Wayback Machine⁹ to provide the version closest to when the article was created or modified. The final HTMLs were saved from the

⁸<https://www.wikihow.com/api.php>

⁹<https://archive.org/about/>

snapshot versions¹⁰. To simplify HTML source code of downloaded related documents, we used ReadabiliPy, a Python wrapper for Mozilla’s Readability.js¹¹ package¹². We then processed HTMLs in two ways:

1. by running it through html2text¹³ and saving the output in Markdown¹⁴ format; or
2. by utilising newspaper3k¹⁵ to extract text content directly.

The former preserves formatting information useful when working with tables and lists, while the latter produces raw text that is simpler to use and analyse. Additionally, we split each Markdown file into a list of passages, and used library-link to merge all lists and tables into one passage. All titles were treated as separate passages to filter them out if necessary.

Fig. 6.5 illustrates a typical human-written summary on WikiHow. Often, target answer summaries end with a sentence encouraging a reader to study the whole article, such as “To learn more, such as how to wash unfinished leather or use a washing machine, keep reading the article!”. To get a shorter, more answer-like versions of summaries, we filtered out these sentences using a set of patterns.

Finally, we rejected articles with missing HTML snapshot links, empty HTML, or empty extracted text. Each instance in our new dataset consists of a question, a human-written article summary treated as the target answer, and a list of parsed texts from the relevant HTML documents cited by the article’s author. Fig. 6.3 shows a simplified example from WIKIHowQA.

Avoiding Overlap in Train-Test

When studying state-of-the-art model performance on ELI5 [59], Krishna et al. [106] observed little to no evidence that analyzed models grounded answer generation in the retrieved documents. They attributed this to a significant overlap (around 43.6%) in the training, validation, and test sets of ELI5. Similarly, Lewis et al. [122] identified train-test overlap in various QA datasets. To avoid this issue in our new resource, we explicitly split questions in a way that no same-topic or paraphrased questions ended up in different splits. To do so, we clustered all questions prior to splitting and assigned all questions belonging to the same cluster to the training split. HDBSCAN¹⁶ [30] was used as the clustering algorithm due to its robustness to noise. Questions were embedded using Universal Sentence

¹⁰<https://archive.readme.io/docs/website-snapshots>

¹¹<https://github.com/mozilla/readability>

¹²<https://github.com/alan-turing-institute/ReadabiliPy>

¹³<https://github.com/aaronsw/html2text>

¹⁴<https://daringfireball.net/projects/markdown>

¹⁵<https://newspaper.readthedocs.io/en/latest/>

¹⁶https://hdbSCAN.readthedocs.io/en/latest/how_hdbSCAN_works.html

The screenshot shows a web page from WikiHow titled "How to Wash Leather". The page includes a sidebar with navigation links like "METHODS", "OTHER SECTIONS", and "Article Summary". The main content area features an "Article summary" section with a green header and a purple-bordered box containing a paragraph of text. Below this, two numbered steps are listed under "Method 1". A third step is listed under "Method 2".

CLEANING ANIMAL MATERIALS » CLEANING LEATHER

How to Wash Leather

Co-authored by Susan Stocker

Last Updated: August 10, 2021 References

Article summary

Although you should only wash leather when it's really dirty, you can use castile soap and water or a specially-formulated leather soap to deep-clean your leather items. Mix a small amount of the soap into a bowl of distilled water, then dip a soft, lint-free cloth in the cleaning solution. Wipe the leather with the cloth, following the grain of the leather as you clean it. Rub using a light, circular motion to remove caked-on dirt or a tough stain. Use a clean, damp rag to remove any soap film and remaining dirt from your leather, then let the item air dry. To learn more, such as how to wash unfinished leather or use a washing machine, keep reading the article!

1 Hand wash leather for basic maintenance and deep cleaning. Hand cleaning is great for spot treating scuff dirt, but it's also an effective way of deep cleaning leather.^[1] Although if your item is expensive or made of hard professional cleaner might be the safest cleaning method.

2 Make a soapy water solution with castile soap. Pour a small amount of castile soap into a bowl of distilled water. Agitate the solution with your hand or ...

Method 2

Using a Washing Machine

1 Pick an item that isn't too expensive. It's important that you know from the outset: there's no guarantee that washing your item won't do damage to it. Durable items, like boots or jackets, might be the best candidates for washing.^[10]

Figure 6.5: Human-written article summary on WikiHow.

Encoder (USE¹⁷) [117]. Parameters “min_cluster_size” and “min_samples” were set to 2 for HDBSCAN, with the default values used for other parameters. We only assign clusters to the training split so that the validation and test splits will have more varied questions, to avoid a potential evaluation bias due to the presence of many similar questions.

6.2.2 Dataset Statistics

The WIKIHowQA dataset include a diverse range of topics, 19 in total, with the most frequent being “health”, “home and garden”, “pets and animals”, and “computer and electronics”. Fig. 6.6 presents the distribution of categories and subcategories in a stacked bar chart. Each bar represents a main category, with its corresponding subcategories shown as segments stacked within the bar. The height of each bar indicates the total number of questions belonging to the category, and the size of each segment reflects the relative contribution of its subcategory. For visibility, only subcategories that constitute more than 1% of the total questions are demonstrated and labeled.

¹⁷, <https://tfhub.dev/google/universal-sentence-encoder-multilingual/3>,

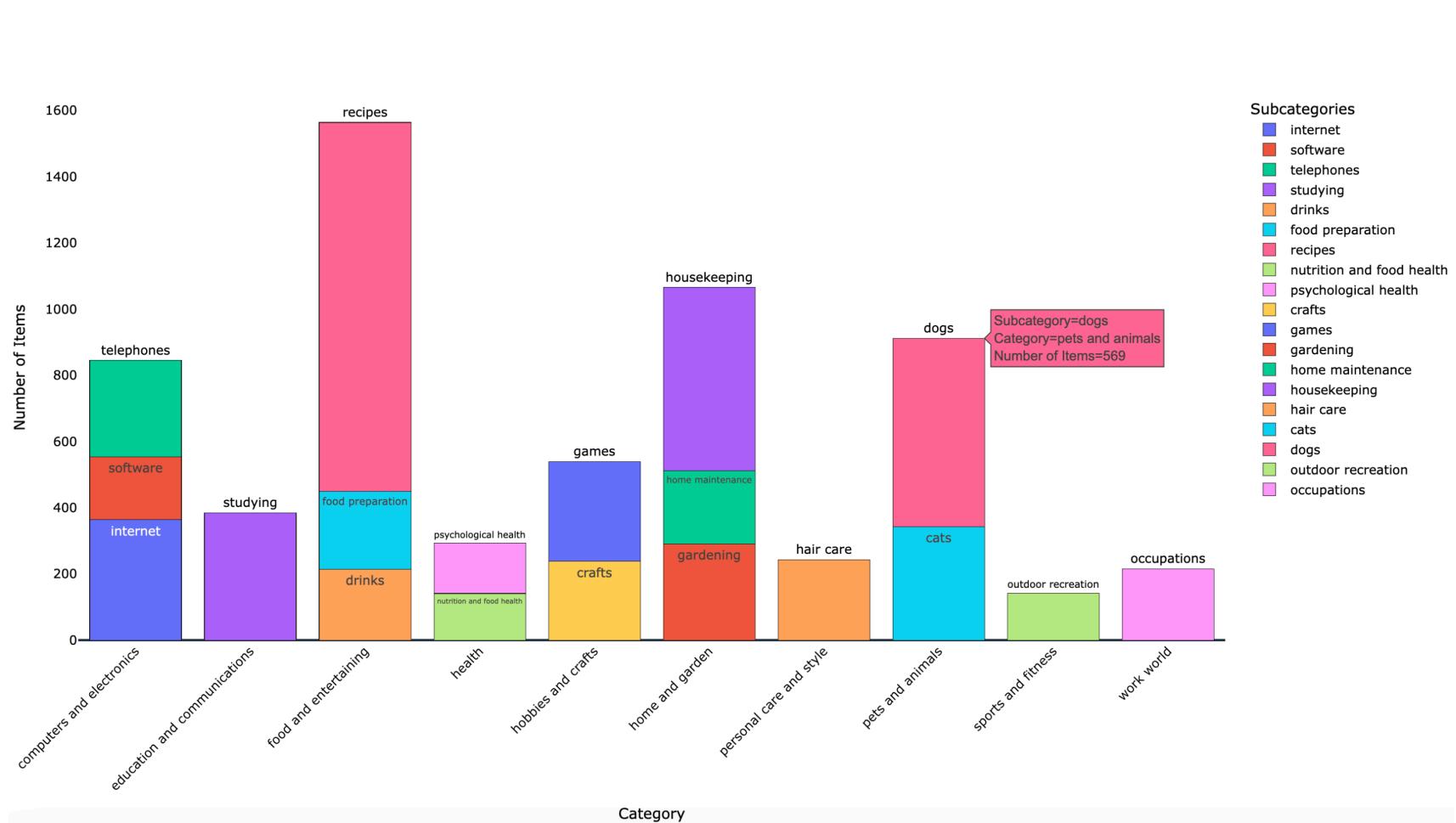


Figure 6.6: Category distribution of WikiHowQA.

Table 6.1: Comparison of WikiHowQA with other QF-MDS and LFQA datasets.

Dataset	# questions	Splits			Reference sources			Answer	
		train (#clusterised/#clusters)	valid	test	# docs	# words	# sents	# words	# sents
WIKIHowQA	11,746	8,235 (2,449/7,272)	1,178	2,333	6.3	1,053.6	65.2	113.05	4.9
AQUAMUSE [110]	5,519	4,555	440	524	6	1,597.1	66.4	105.9	3.8
ELI5 [59]	272,000	237,000	10,000	25,000	–	857.6	–	130.6	–
HowSUMM-METHOD [20]	11,121	8,856	1,122	1,143	11.19	1,455.52	71	539.11	31.33
HowSUMM-STEP [20]	84,348	67,403	8,248	8,697	9.98	1,357.37	66.47	98.98	5.23

Table 6.1 presents a comparison of WIKIHowQA with other query-focused multi-document summarization and long-form question-answering datasets. ELI5 [59] represents the largest dataset in terms of the number of questions but features the lowest average word count in reference documents. It is important to note that the reference document in this dataset is an aggregation of top- k retrieved passages, which explains the absence of some statistics in Table 6.1. In contrast, WIKIHowQA ensures a higher word count per document. This implies a more in-depth coverage per topic, which is essential for training QA systems to generate more detailed and informative answers. Additionally, the number of answer words in WIKIHowQA is competitive, offering substantial content that can challenge and thereby improve the capabilities of QA models.

WIKIHowQA includes a significant number of questions, with totals that are twice as many as those in AQUAMUSE and comparable to HowSUMM-METHOD. Furthermore, each question in WIKIHowQA is supported by multiple documents, averaging 6.3 documents per question. The number of supporting documents in WIKIHowQA compares favorably with other datasets. While it does not reach the high of 11 documents per instance seen in HowSUMM-METHOD, the requirement for such a high number of documents in HowSUMM-METHOD is a design choice aimed at supporting the creation of exceptionally detailed answers. This results in an average answer length of more than 539 words, which is about five times longer than those found in other datasets. This approach, while thorough, highlights its specialized use for tasks requiring in-depth responses, differing significantly from datasets like WIKIHowQA that focus on more concise and user-direct answers.

While our developed dataset WIKIHowQA shares some common statistical characteristics of other datasets, the construction process, source-target mappings and task applications differ as discussed in Section 2.

6.2.3 Quality Verification

Since the relevant passages in our WikiHowQA datasets are sourced from web pages cited by the WikiHow article author, they may not contain the exact words or phrases from target

Table 6.2: Percentage of novel n-grams.

Dataset	% novel n-grams		
	uni-grams	bi-grams	tri-grams
WIKIHowQA	11.75	54.87	85.02
DUC 03–04	27.74	72.87	90.61
HowSUMM-METHOD	15.20	52.70	81.90
HowSUMM-STEP	9.80	47.30	78.90

answers. Therefore, to verify the quality of our benchmark, we assess the feasibility of constructing an answer from the given supporting documents. As this is a crucial aspect for a multi-document non-factoid question answering resource, we evaluate our benchmark using both automatic metrics and thorough human evaluation.

Automatic Metrics

The relevance of the supporting documents in our WIKIHowQA dataset is first evaluated using the Novel N-Gram Percentage metric [166] as a measure of word intersection between answers and documents. This metric, commonly used for summarization datasets, allows for a strong upper bound comparison. Results presented in Table 6.2 show that the passage answers in our dataset correspond to the content of the supporting documents with a relatively high n-gram percentage score, similar to that of HowSUMM-METHOD. We also include results for the Document Understanding Conference (DUC) dataset as an upper bound as it is a high quality human crafted summarization dataset (as reported by Fabbri et al.).

Secondly, we report the average coverage, density and compression metrics for WIKIHowQA [73]. These metrics are commonly used to characterize the quality and difficulty of summarization tasks. We follow Fabbri et al. [56] who adapted these measures for the multi-document summarization use case. For the WIKIHowQA, the scores are notably high, with coverage at 0.89, density at 1.86, and compression at 81.46, indicating a robust alignment between source documents and summaries. It is important to note that Tejaswin et al. [184] and Bommasani and Cardie [19] propose additional metrics to compare datasets, some of which are a reformulation of those presented by Grusky et al. [73].

The above metrics give a general indication of the dataset quality, although even when examples from our dataset are deemed lower quality, for example when they have low semantic similarity, we find that there are instances where an answer can typically still be constructed from the supporting documents. Fig. 6.3 demonstrates that answers can be successfully constructed from documents even if they have low semantic similarity. For instance, in the first sentence of an answer for the question “How to cook pork tenderloin”

(Fig. 6.3), although the target length of 55 minutes was not mentioned in any document, it can be inferred by adding up roasting time of 30 minutes on one side and 25 minutes on the other side. Similarly, the suggestion “flipping it occasionally” in the last sentence can be rephrased from “turning every 1 1/2 to 2 minutes”.

Human Evaluation

To verify our observations and ensure the quality of our dataset, we conducted a crowdsourcing study with the goal of verifying the feasibility of answer construction. All our crowdsourcing studies within this research were reviewed and approved by the Human Research Ethics Committee of RMIT University¹⁸. We use Surge AI¹⁹ as our data labeling platform, which provides a workforce of highly skilled and educated native speakers, ensuring high-quality data labeling at scale, allowing for higher quality labeling compared to traditional platforms such as Mechanical Turk.

Each participant task in the study consisted of a QA pair, one supporting document, and three evaluation components. The first evaluation required annotators to classify the question into the appropriate category from a provided list. This step served as an attentiveness check. Annotators were presented with four categories in a random sequence, one of which was the original category derived from the WikiHow website, while the remaining three were randomly selected from the other eighteen categories. This task was designed to ensure that the annotator was paying attention and had read the question carefully. The tasks were rejected if the classification was incorrect. The second evaluation required annotators to read the answer and the supporting document, and use a provided highlighting tool to mark any words or phrases in the supporting document that were directly relevant to the given answer, or could be used to reason or conclude it. In the final step of the process, annotators evaluated the relevance of the provided documents to the answer on a five-point Likert scale (from 0, “Very Irrelevant”, to 4, “Very Relevant”). In total, 31 randomly selected QA pairs and their associated supporting documents were annotated, resulting in 104 participant tasks, with only 4 being rejected. Each assessment was carried out by three workers, receiving compensation of \$0.7 per completed task. The study involved 22 fluent English speakers.

Figure 6.7 presents the interface used to assess the feasibility of constructing answers from the provided supporting documents in our WIKIHowQA. The example assessment shown in the figure is a random selection from the validation data collected from the workers. The detailed instructions are shown in Figure 6.7.

¹⁸Reviewed and approved by Research Ethics Committee of RMIT University according to the IRB procedures for research involving human subjects. The IRB approval number is 66-19/22334.

¹⁹<https://www.surgehq.ai/faq>

6. WikiHowQA: A COMPREHENSIVE BENCHMARK FOR MULTI-DOCUMENT NON-FACTOID QUESTION ANSWERING

Please, read the following question, answer and the supporting document:

“Question: How To Book a Flight? ”

Answer: To book a flight online, you can either go directly to a certain airline's website, like JetBlue.com, or visit website: Follow the website's instructions for choosing your date and airport, and consider clicking the "My dates are flexible" but option. Then, follow the website prompts and include the necessary info, like your name and credit card number to purch

Please select the category that best fits the question from the list provided.

- Travel
- Personal Care and Style
- Food and Entertaining
- Pets and Animals

Please, reread the question and answer carefully.

Use the highlighting tool below to mark any words or phrases in the given supporting document that are directly related to the given answer or to example, if the answer is "fry for 55 minutes", you may highlight two spans "first fry for 25 minutes", "then turn", and "fry for another 30 minutes" ir evaluate the quality of the supporting documents and determine whether the answer can be accurately constructed from the supporting dc highlight specific bits of text rather than entire sentences or paragraphs and to aim for other participants to use the highlighted parts of the seeing the given answer.

Thank you for your careful attention to this task.

Highlight the text to create a new label. Click on an existing label to remove it.

Relevant

How to avoid every common mistake when booking a flight — Quartz

As the Travel Editor for CBS News, people expect that I spend weeks, even months, researching the process and logistics of travel. And I do. But what about airfares? I only spend minutes. It's not just that I understand what makes a reasonable or crazy fare for each route. I also know I There is a science to airfare: RELEVANT The day, time of day, and the time of the month that you book matter. RELEVANT The number one mistake I see most travelers make is to book too early RELEVANT . Unless you are planning travel for high-traffic days, like Christ possible fare 45 days out for domestic travel and 60 days out for international RELEVANT .

Outside of that 45-day window RELEVANT most airline computers aren't programmed to give you any kind of a deal. It's all about computer mode flew on Flight 405 to Cleveland last February and in February of 2012 as well. They make their projections of the load for this February based on !

If possible, I aim to book flights after the 7th of every month RELEVANT , as booking is busier right after payday on the first and 15th.

Whenever possible, I shop for fares on one computer and then clear my cookies RELEVANT or book on another computer altogether. If you want there's no proof or hard evidence , there is growing anecdotal evidence that suggests airlines are tracking your behavior online RELEVANT , and pc up for flights that are searched frequently.

This post is part of a series in which LinkedIn Influencers share their secrets to being more productive. See all their #productivityhacks here.

Please, assess the relevance of the given supporting document to the answer.

- Very Irrelevant ?
- Somewhat Irrelevant ?
- Neutral/Unsure ?
- Somewhat Relevant ?
- Very Relevant ?

Figure 6.7: Interface for the assessment of supporting document relevance (cropped from the right to fit the page for better visibility).

The short task description presented to the workers was as follows: "Welcome to our study! We are interested in understanding how well "supporting documents" can be used to answer a question. Thank you for participating in the study. Please review the participant information form before proceeding." Fig. 6.8 illustrates a randomly selected result of this annotation process, as captured in the (simplified) annotation interface used in the study.

The results showed that the workers deem the supporting documents to be mostly "Somewhat Relevant" in relation to the given answers, with a mean document relevance score of 2.59. The inner-annotator agreement is moderate, with a Cohen's Kappa [10] value of 0.51, indicating a reasonable consistency in the annotators' assessments of document relevance. In terms of related text selections within the documents, the mean overlap score between assessor pairs was 0.65, calculated using the overlap coefficient [193]:

$$\text{Overlap}(H_1, H_2) = \frac{|H_1 \cap H_2|}{\min(|H_1|, |H_2|)}$$

where H_1 and H_2 are the sets of unique words from two annotations. This is consistent with previous similar research [152] and the results from Section 3.3.4, and indicates that the

"Question: How To Park on a Hill? "

Answer: To park on a hill, start by pulling up to the space you want to park in so you're parallel with the curb. Then, turn your wheels away from the curb if you're parking uphill or toward the curb if you're parking downhill so your car doesn't roll into traffic if the brakes fail. Next, let your car roll downhill slightly until you feel the front tire touch the curb. Once the front tire is touching the curb, put your car in park and engage the emergency brake.

Highlight the text to create a new label. Click on an existing label to remove it.

Relevant

Parking on a Hill - What to do if You're Uphill or Downhill

Uphill: When headed uphill at a curb, turn the front wheels away from the curb RELEVANT and let your vehicle roll backwards slowly until the rear part of the front wheel rests against the curb RELEVANT using it as a block.

Downhill: When you stop your car headed downhill, turn your front wheels toward the curb RELEVANT . Let your vehicle roll slowly until the front tire rests against the curb using it as a block RELEVANT .

When headed uphill (or downhill) with no curb, always turn your front wheels to the right so that if your vehicle moves, it will roll off the highway, not into traffic.

Please, assess the relevance of the given supporting document to the answer.

- Very Irrelevant ?
- Somewhat Irrelevant ?
- Neutral/Unsure ?
- Somewhat Relevant ?
- Very Relevant ?

Figure 6.8: WIKIHowQA quality annotation.

annotators had good agreement with each other on the parts of the documents that could be used for associated answer construction. Overall, these findings demonstrate that the provided documents are relevant and contain the necessary information for the answers.

6.3 Baseline Models

This section describes the baseline models for the WIKIHowQA. The task involves generating an answer to a non-factoid question that is grounded in a set of relevant documents. These documents provide the necessary information from which the answer can be reasoned or inferred.

As a reference point for performance evaluation, we use the Retrieval-Augmented Generation (RAG) [121] architecture, which is a common choice for abstractive summarization and QA tasks [59]. Since our dataset provides a small predefined set of relevant supporting documents for each question, we use a retriever to filter out passages from those documents that are less relevant to the question, to help fit more relevant passages into the limited input of a generator. Then, a generator is tasked with generating an answer conditioned on input passages. We experimented with DPR [149] as the retriever and BART-large [120] as the generator.

When building model input, we first ranked passages from supporting documents based on their relevance to the question using DPR, then truncated them to retain the maximum number of complete passages in the ranking that fit within the maximum input length of BART. Passages were then reordered based on their source documents, and combined into a single input string of the format “*<question></q><d><p>doc 1 passage 1</p><p>...</p></d><d><p>doc 2 passage 1</p>...</d>*”. Finally, BART is trained to generate gold answers conditioned on this context. Table 6.3 show significant hyper-parameters used to train the BART model and Table 6.4 contains the hyper-parameters for decoding. The number of training epochs was controlled through early stopping based on the validation loss (typically 2-4 epochs).

To establish a stronger baseline, we also evaluated the performance of the GPT-3 [26] variant *text-davinci-003*²⁰ in a zero-shot setting. As a model at least 400x times larger than BART, it forms an upper bound for model performance. We compared two prompt kinds for GPT-3, one with the question and another with the question and relevant passages retrieved via DPR.

Standard automatic evaluation metrics for abstractive QA, Rouge-X and BertScore, are reported in Table 6.5. However, these scores are known to poorly reflect actual model performance [49, 106]. While recent QA-based evaluation metrics for summarisation

²⁰<https://beta.openai.com/docs/models/gpt-3>

Hyper-parameter	Value	Hyper-parameter	Value
learning rate	5e-05	repetition_penalty	5.0
train batch size	2	top_k	10
eval batch size	4	top_p	0.95
seed	42	temperature	1.2
gradient accumulation steps	16	no_repeat_ngram_size	2
total train batch size	32		
optimizer (Adam) betas	$\beta_1 = 0.9, \beta_2 = 0.999$		
optimizer (Adam) epsilon	1e-08		
lr scheduler type	linear		
number of epochs	2-4 (early stopping)		

Table 6.3: Training hyper-parameters for BART.

Table 6.4:
Decoding hyper-parameters for BART.

Table 6.5: Baseline models for WIKIHowQA.

Model	Automatic Evaluation				A/B Human Evaluation			
	Rouge-1	Rouge-2	Rouge-L	BertScore	100 instances	Prefer Model	Prefer Gold	Tie
DPR + BART	39.8	12.4	23.0	0.881	13	52	35	
<i>text-davinci-003</i>	32.2	8.5	19.7	0.873	18	53	29	
DPR + <i>text-davinci-003</i>	35.4	9.2	20.2	0.868	56	15	29	

All differences are statistically significant. (Student t-test, p-value < 0.01).

seem to better correlate with human judgements [50, 165], their adaptation to NFQA is not straightforward and remains an important area of research, as answers in NFQA are expected to contain facts not mentioned in supporting documents. Instead, we report the results obtained through our human evaluation framework as a more accurate measure of model performance (see Section 6.4.2).

6.4 How To Evaluate How-To Answers

To enhance the reliability of multi-document non-factoid question answering evaluations, we introduce a streamlined human evaluation framework. This framework is utilized to assess the performance of baseline models within our benchmark.

Table 6.6: Examples of deterioration types.

Deterioration Type	Original Sentence	Deteriorated Sentence
Number Swap	Sauté the onions for 5 minutes.	Sauté the onions for 15 minutes.
Sentence Negation	Turn the vehicle off and open the hood.	Don't turn the vehicle off and open the hood.
Antonyms Swap	Adjust your iron to hot for linen.	Adjust your iron to cold for linen.
Entity Swap	As a rabbi, you'll train in a branch of Judaism.	As a rabbi, you'll train in a branch of Christianity.

6.4.1 Evaluating Human Evaluation

Initially, we employed a conventional human evaluation approach, as described in previous studies [59, 106], where assessors are presented with both model-generated and gold standard answers in a randomized order and asked to select their preferred response. However, our findings indicated that, even when presented with gold standard questions and evaluated by high-performing assessors, simpler model-generated answers were frequently (48% of the time) preferred over reference answers, despite containing factual inaccuracies. Manual inspection revealed that choosing between two well-formulated answers was challenging for participants, in line with previous research of Clark et al., Krishna et al.. To quantify the ability of annotators to identify factual inconsistencies in answers, we conducted a crowdsourcing evaluation experiment. For this experiment, we intentionally introduced factual errors into a selection of answers from the WIKIHowQA, creating a controlled set of deteriorated answer versions.

Answer Deteriorations

To assess the accuracy of answers and the effectiveness of human evaluation methods, we introduced specific types of factual inaccuracies into a subset of answers. This approach is inspired by methodologies used by Kryscinski et al. [108] for evaluating the factual consistency of abstractive text summarization models. We implemented the following types of errors, with illustrative examples provided in Table 6.6:

- **Number Swap:** Replacing a numerical value in the original sentence with a different value. For example, the sauté time was changed from 5 minutes to 15 minutes.
- **Sentence Negation:** Rephrasing the original sentence using negations and different grammar structures, altering the meaning and making the answer factually inconsistent. For instance, the instruction to cool the pouch before removal was negated.
- **Antonym Swap:** Replacing an adjective in the original sentence with its antonym. For example, the heat setting on the iron was changed from hot to cold.
- **Entity Swap:** Replacing a proper noun or named entity in the original sentence with a different entity, altering the context and making the answer factually inconsistent. For example, changing "Judaism" to "Christianity".

Note that while the above methods were employed to generate a large set of incorrect answers, they did not always yield grammatically or semantically correct results. Therefore, each generated answer underwent manual review and editing to ensure that they remained

contextually plausible and grammatically correct. For all types of deteriorations — such as identifying adjectives, entities, numbers, auxiliary verbs, and lexical verbs — we utilized the Spacy NLP library²¹.

To construct the evaluation set, we randomly selected five QA pairs from each of the nineteen question categories available on WikiHow. Subsequently, we introduced various types of inaccuracies into four out of the five answers per category, employing the methods previously outlined. This approach resulted in an average of 4.5 modifications per answer. After incorporating these inaccuracies, each deteriorated answer underwent a manual review process to ensure that it remained contextually relevant and grammatically correct. The specific steps in the editing process included:

- **Number Swap** involved randomly replacing digits within the text. However, during the review phase, these numbers were carefully adjusted to ensure they remained plausible within the context yet introduced a factual error that was subtle yet discernible.
- In **Sentence Negation**, lexical verbs were modified by preceding them with "don't" or "doesn't," and auxiliary verbs were replaced with their negative forms. For complex sentences with multiple clauses, this method sometimes led to grammatical inconsistencies. Where necessary, entire sentences were slightly modified to accommodate the negation while maintaining grammatical correctness.
- For **Antonym Swap**, antonyms were sourced using the WordNet Lexical database [132]. The most contextually appropriate antonym was chosen to replace the original adjective in the sentence. Adjustments were made during manual review if the initially chosen antonym did not fit the sentence context appropriately.
- **Entity Swap:** At times, this deterioration involved manual intervention to identify and replace named entities when the automatic methods did not suffice. The substituted entities were carefully selected to ensure that answers remained contextually coherent but factually incorrect.

Using the newly created set, we conducted two separate evaluation rounds involving a total of thirty-four crowdsourcing workers, who assessed all ninety-five QA pairs. Importantly, no worker participated in both rounds, ensuring an unbiased evaluation process.

First Evaluation Round

The first trial aimed to establish the baseline performance of the standard evaluation framework, in which assessors evaluate answers without the assistance of any on-screen

²¹<https://spacy.io/models/en>

6. WikiHowQA: A COMPREHENSIVE BENCHMARK FOR MULTI-DOCUMENT NON-FACTOID QUESTION ANSWERING

Question: How To Stop Breastfeeding a Toddler

First, select the appropriate category for the question and rate how familiar you are with the question topic.
Next, evaluate the usefulness of the answer given below and identify useful and misleading text spans in the answer.

Please select the category that best fits the question from the list provided.

- Philosophy and Religion
- Arts and Entertainment
- Family Life
- Food and Entertaining

Please select the option that best reflects your familiarity with the topic of the question.

- Very Familiar ?
- Somewhat Familiar ?
- Neutral ?
- Somewhat Unfamiliar ?
- Very Unfamiliar ?

Answer: To stop breastfeeding a toddler, start by eliminating 3 nursing sessions a day for a period of 1 week. Then, eliminate 4 sessions a day for the next week. This will gradually wean your toddler off of nursing. During the weaning process, try to avoid triggers that remind your baby of nursing, like sitting in the chair that you normally nurse. Additionally, if your toddler asks to nurse, attempt to distract your child by going for a walk together or singing a song.

Please, rate the usefulness of the answer above.

Please select the option that best reflects the usefulness of the answer provided. By "usefulness", we mean how well the answer addresses the question. If you are unable to evaluate the answer, please select "I don't know".

- Very useful ?
- Somewhat useful ?
- Neutral ?
- Somewhat useless ?
- Very useless ?
- I don't know ?

Please, reread the question and answer carefully.

Use the highlighting tool below to mark any words or phrases in the answer that are useful or misleading.

To select a span type, click on the "useful span" or "misleading span" option. Then, select the words or phrases in the answer that you want to highlight. You must highlight at least one useful or misleading span. You can also highlight several useful and misleading spans in one answer.

Note: It is generally not recommended to highlight the entire answer, even if it is very useful. Instead, try to highlight specific words or phrases that are particularly useful. However, suppose the entire answer is very useful, and no specific words or phrases stand out as particularly useful. In that case, you may choose to highlight the entire answer. Your highlighting will be used to assess the completeness and accuracy of the answer and the relevance of the information provided.

Thank you for your careful attention to this task.

Highlight the text to create a new label. Click on an existing label to remove it.

Select multiple tags? ?

Useful span Misleading span

To stop breastfeeding a toddler, start by eliminating 3 nursing sessions a day for a period of 1 week. USEFUL SPAN Then, eliminate 4 sessions a day for the next week. This will gradually wean your toddler off of nursing. During the weaning process, try to avoid triggers USEFUL SPAN that remind your baby of nursing, like sitting in the chair that you normally nurse. attempt to distract your child USEFUL SPAN by going for a walk together or singing a song.

Figure 6.9: Interface for the first trial of the factual inconsistency identification study (cropped from the right to fit the page for better visibility).

relevant information. Each HIT included one QA pair where the answer may have been deteriorated. Similar to Section 6.2.3, HITs began with the attentiveness test of question category classification. Assessors then indicated their familiarity with the question and evaluated the usefulness of the answer on a five-point Likert scale (from “Very Unfamiliar” / “Very Useless” to “Very Familiar” / “Very Useful”).

Finally, they were tasked with highlighting spans within the answer that they found either useful or misleading, thereby providing a more nuanced understanding of the answer’s perceived value. Figure 6.9 illustrates an interface of the first trial run for the study with a random annotation example which happens to have an example where an assessor failed to find any deterioration.

Second Evaluation Round

The second trial featured the same evaluation tasks, except this time assessors were provided with highlighted relevant passages for each sentence in the answer, accessible by clicking on the sentence. Top-ranked passages according to DPR (see Section 6.3) were selected for each sentence, ensuring the overall length is under 5000 characters to fit on the screen. Passages were displayed in their original order, grouped by source documents.

Passage highlighting came from the model developed in Section 3.5.1, designed to simulate user gaze during non-factoid question-answering evaluation. Assessors could freely examine the passages while determining the usefulness of the answer and selecting useful or misleading spans in the answer. The choice of the presentation form was based on several factors.

Firstly, it allowed for a more interactive and engaging experience for the assessors, enabling them to delve deeper into the context of each sentence. Secondly, it provided a clear visual cue to the assessors about the relevance of each sentence, thereby facilitating a more accurate and efficient evaluation process. Given that assessors often face challenges due to their unfamiliarity with the topic or potential distractions, we aimed to enhance their ability to assess with higher accuracy by providing relevant passages as supporting information. This approach, designed to be simple yet effective, prioritized the provision of relevant information over more elaborate feedback mechanisms like clicks or gaze tracking, which are potential areas for future research.

After comparing different forms of passage presentation through trial runs and discussions among the authors, we found that the form offering easy accessibility by clicking on the sentence was the most effective. Figure 6.10 shows an interface of the second trial run of the study with a random annotation example.

Inconsistency Detection Between Rounds

As a measure of the ability to spot factual inconsistencies, we calculated the inconsistency detection rate (IDR), defined as the average percentage of identified deteriorations. A deterioration was marked as identified if it was highlighted as part of a misleading span.

The IDR was calculated for each round, and the results showed a statistically significant improvement from the first to the second trial (Student's t-test, $t = 2.57, p < 0.01$), with IDR scores increasing from 0.21 in the first trial to 0.35 in the second.

Additionally, we monitored the occurrence of false positives — instances where participants incorrectly identified an answer as inaccurate. The first trial resulted in four false positives, while the second had only two, indicating a better discernment in the latter.

Figure 6.11 presents the percentage of recognized deteriorations across various WikiHow question categories, comparing the two trials. Most categories demonstrate an increase in detection in the second trial, with the exception of Education and Communication, Family Life, Finance and Business, Philosophy and Religion, Relationships and Youth. These categories generally involve topics that are subjective and open-ended, which may complicate the detection of factual inaccuracies.

While question familiarity scores were consistent between the two trials, IDR improvement was persistent and statistically significant (Student's t-test $t - statistic = 2.18, p < 0.05$) in the second trial for both high and low familiarity. This suggests that *participants were able to assess more effectively with an access to relevant information sources, regardless of their familiarity with a particular question.*

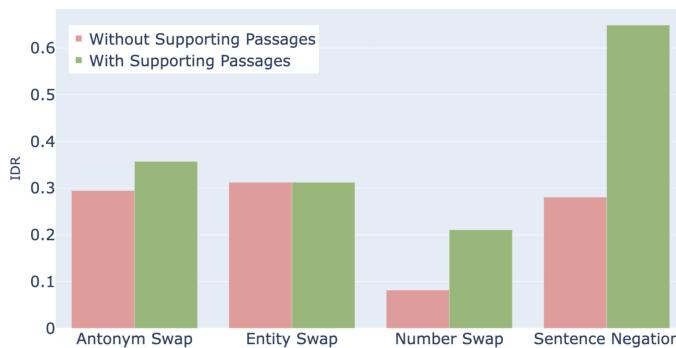


Figure 6.12: IDR by type.

As shown in Fig. 6.12, there was a general improvement in detection of almost all deterioration types in the second trial. Notably, the detection rate for sentence negations and number swaps increased by 131% and 158%, respectively. These increases were statistically significant, as confirmed by Student's t-tests ($t - statistic = 4.2, p < 0.01$ for sentence negations; $t - statistic = 3.18, p < 0.01$ for number swaps).

These findings underscore the importance of a well-structured evaluation framework

Question: How To Strip Cloth Diapers

First, select the appropriate category for the question and rate how familiar you are with the question topic.
Next, evaluate the usefulness of the answer given below and identify useful and misleading text spans in the answer.

Please select the category that best fits the question from the list provided.

- Sports and Fitness
- Philosophy and Religion
- Pets and Animals
- Family Life

Please select the option that best reflects your familiarity with the topic of the question.

- Very Familiar ?
- Somewhat Familiar ?
- Neutral ?
- Somewhat Unfamiliar ?
- Very Unfamiliar ?

Please carefully read the provided answer below and rate the usefulness of the answer above.

If you are NOT SURE about ANY facts mentioned, please consult the supporting passages (accessible by clicking on each sentence) that are related to the question. It's important to use these passages as they will help you make sure the answer is correct and relevant. Click on each sentence to access the passages and read them. We have highlighted the most important parts of the passages for you, but make sure to read through the entire passage as other parts may also be useful.

Answer:

Before you strip cloth diapers, wash and dry them like you normally would so they're clean.

► You can also make your own by mixing equal parts washing soda, borax, and Calgon.

▼ Next, soak the diapers in the mixture for around 6 hours, stirring them around occasionally to help release the minerals in the fabric.

4.) Soak all your items in the water until it cools (at least 2 hours) or overnight (but no more than 8 hours). Ideally, you will soak between the minerals from the fabrics.

...
6.) After the water is cool drain the tub, and squeeze all water from the items. 7.) Complete with a WATER ONLY wash cycle. This will n

Boil them on the stove for 5-10 minutes. Find the biggest pot possible, fill with water and bring to a boil. Put your inserts or diapers in t

► Finally, remove the diapers, and rinse them in cold water before hanging them up to air dry.

Please select the option that best reflects the usefulness of the answer provided. By "usefulness", we mean how well the answer addresses your needs.

If you are unable to evaluate the answer, please select "I don't know".

- Very useful ?
- Somewhat useful ?
- Neutral ?
- Somewhat useless ?
- Very useless ?
- I don't know ?

Please, reread the question and answer carefully.

Use the highlighting tool below to mark any words or phrases in the answer that are useful or misleading. To select a span type, click on the "useful span" or "misleading span" option. Then, select the words or phrases in the answer you want to highlight. You can also highlight several useful and misleading spans in one answer.

It is important to study the supporting passages by clicking on each sentence in the answer above. These passages may contain additional information that can help you identify useful or misleading spans in the answer.

You must highlight at least one useful or misleading span.

Highlight the text to create a new label. Click on an existing label to remove it.

Useful span

Misleading span

Figure 6.10: Interface for the second trial of the factual inconsistency identification study (cropped to fit the page for better visibility).



Figure 6.11: IDR by categories.

that not only challenges the participants’ detection capabilities but also supports them with the necessary resources to perform the task effectively. The improvements across the trials highlight the potential for further refining the human evaluation process in non-factoid question answering tasks.

6.4.2 How Good Are The Baselines, Really?

To assess the effectiveness of the lower- and upper-bound baseline models introduced in Section 6.3, we employed our human evaluation framework. Specifically, we compared models through A/B testing [106] while incorporating highlighted relevant passages as reference sources of information.

Each evaluation session began with an attentiveness task to classify the question category, as described in Section 6.2.3, ensuring that assessors were engaged and focused. Subsequently, they were presented with a pair of answers — one gold and one model-generated — in a randomized order. Assessors were instructed to evaluate the usefulness of each answer and select the one they found more useful, without knowing which was the gold standard and which was the model-generated answer. We provided ranked relevant passages for each sentence in each answer separately, accessible when clicking on a sentence. This allowed participants to make more informed decisions about factual consistency, as demonstrated in the previous section.

In this experiment, twenty-eight workers evaluated gold answers to one hundred random test questions against answers generated by DPR + BART, *text-davinci-003*, and DPR + *text-davinci-003*. To eliminate potential bias, workers did not evaluate the same question twice. Each HIT was compensated with \$0.50. Figure 6.13 and Figure 6.14 demonstrate the interface used for A/B evaluation of our lower and upper-bound models, using a random annotation example for DPR + *text-davinci-003* model evaluation. In the example, the top part of Figure 6.13 shows Answer A, which is generated by the model, while the bottom part of Figure 6.14 contains Answer B from the dataset.

"Question: How To Make a Tortillon?"

First, select the appropriate category for the question.

Please select the category that best fits the question from the list provided.

- Philosophy and Religion
- Education and Communication
- Pets and Animals
- Hobbies And Crafts

Please carefully read the provided ANSWER A below and rate its usefulness.

If you are NOT SURE about ANY facts mentioned, please, consult the supporting passages (accessible by clicking on each sentence) that are related to the question. It's important to use these passages as they will help you make sure the answer is correct and relevant. Click on each sentence to access the passages and read them. We have highlighted the most important parts of the passages for you, but make sure to read through the entire passage as other parts may also be useful.

Answer:



Making a tortillon is simple and can be done with everyday materials.



First, take a sheet of A4 paper, mark 1" down on one side and 1" up on the other; then rule a pencil line between the two points a

| Take a sheet of A4 paper and mark 1" down on one side and 1" up on the other. Rule a pencil line between the two points and the

...

| Once the paper is tightly rolled, hold it firmly and push out the centre with a very thin knitting needle / thick wire or something sim

After you are completed with your essay, insert an A4 Size Paper into the printer. Then go ahead and click "print" on your desktop.



Starting at the narrow end, roll the paper fairly tightly towards the wider edge; then hold it firmly and push out the centre with a w



Secure the paper with a small strip of adhesive tape and trim any excess paper from the top.



Finally, rough up the point with some medium sandpaper to give a softer, velvet texture.



This technique can be used with pencil, pastel, and charcoal to blend and push color around on the drawing paper, softening edges.

Please select the option that best reflects the usefulness of the answer provided. By "usefulness", we mean how well the answer seeks.

If you are unable to evaluate the answer, please select "I don't know".

- Very useful 
- Somewhat useful 
- Neutral 
- Somewhat useless 
- Very useless 
- I don't know 

Figure 6.13: Interface for A/B human evaluation - Top part.

6. WikiHowQA: A COMPREHENSIVE BENCHMARK FOR MULTI-DOCUMENT NON-FACTOID QUESTION ANSWERING

Please carefully read the provided ANSWER B below and rate its usefulness.

If you are NOT SURE about ANY facts mentioned, please consult the supporting passages (accessible by clicking on each sentence) that are related to the question. It's important to use these passages as they will help you make sure the answer is correct and relevant. Click on each sentence to access the passages and read them. We have highlighted the most important parts of the passages for you, but make sure to read through the entire passage as other parts may also be useful.

Answer:

▼

To make a tortillon, start by measuring 1 inch down the side of a piece of copy paper and marking it with a pencil.

Tortillons sometimes collapse on themselves when they're old, first losing the sharp point and then just flattening. You can poke it through again with a bit reserving blunted ones for blending large areas.

...

Also, using a used stump or tortillon for adding color means that you're more likely to use enough texture to fill the grain of the paper and have gentle, soft and you may have to scrub it into a patch of color more than once to get your softly shaded areas worked out just right, but they give immense control. The soft shading. This is great when you're using oil pastels with hard ink lines or other mediums where you don't want them to cover opaquely or have strong i

A tortillon is a tightly rolled sheet of paper, the inside of which is pushed out into a pencil shape and used by pencil, pastel and charcoal artists to blend and edges and graduating tones. You may want to try experimenting with various types of paper, the commercial tortillons that I've come across are made from paper but I've found that just about any paper will do. I make mine from ordinary copy paper straight out of my printer. Very little practice is required to pro

...

| Take a sheet of A4 paper and mark 1" down on one side and 1" up on the other. Rule a pencil line between the two points and then cut along with scissors
I finish by roughing up the point with a piece of medium sandpaper, this gives a softer, velvet texture, especially if using ordinary printer paper as I do. The tortillon. Try to use a clean one for lighter areas of blending and darker, dirtier ones for dark areas such as hair and deep shadows. Use the tortillon at a slight angle. By the way, some nice, soft skintones can be achieved using a soft chamois leather or ordinary kitchen / toilet tissue over the fingertip, especially around the

►

Flip the paper and mark it the same way on the other side, then connect the 2 marks with a straight line.

►

Next, use scissors to cut along the line so you end up with 2 identical pieces of paper.

►

Then, grasp one of the pieces by the narrowest end and roll it tightly towards the outer edge of the paper.

►

Finally, use a piece of thick wire to push the center out so the tip is pencil-shaped, and secure the paper with a small piece of tape.

Please select the option that best reflects the usefulness of the answer provided. By "usefulness", we mean how well the answer addresses the question a seeks.

If you are unable to evaluate the answer, please select "I don't know".

- Very useful [?](#)
- Somewhat useful [?](#)
- Neutral [?](#)
- Somewhat useless [?](#)
- Very useless [?](#)
- I don't know [?](#)

Please, select which answers A or B would provide a more helpful/useful answer overall to a person asking a question.

- A is better
- B is better
- Both good
- Both bad

Figure 6.14: Interface for A/B human evaluation - Bottom part.

Human evaluation results are presented in Table 6.5, with all differences being statistically significant (Student's t-tests, $p < 0.01$). Compared to the initial human evaluation attempt, the simpler DPR + BART model scored lower but was still preferred over gold answers in some instances. Unlike automatic evaluation, human evaluation aligns with

our initial observation that the quality of answers from *text-davinci-003* is much higher than from BART. When prompted with supporting passages, *text-davinci-003* generates significantly better answers based on human judgement (Table 6.5), which highlights the value of reference documents in WIKIHowQA.

While *text-davinci-003* mainly generated grammatically plausible and logically sound answers, our manual analysis revealed that it frequently failed at factual consistency and overall coherency, especially for questions that require a step-by-step instruction. Coupled with the fact that assessors still preferred answers from *text-davinci-003* in some of these cases, the proposed evaluation framework could be further improved to guide assessors even better in their judgement.

The challenge is twofold: first, the provided relevant passages do not always include the necessary information to fact-check the answer; second, the presentation and fluency of model answers can mislead assessors, even when the overall instruction is impossible to follow. This requires assessors to have an additional skill set and reveals room for improvement in the current evaluation framework, highlighting an avenue for future research to develop a more robust system specifically tailored to assessing answers in the *INSTRUCTION* question category.

6.5 Chapter Summary

In this chapter, we have presented a new dataset and benchmark for multi-document non-factoid question answering, WIKIHowQA, sourced from the WikiHow website and consisting of passage-level answers to "how-to" questions. Our released test collection fills a critical gap in currently available resources and evaluation testbeds for multi-document non-factoid QA, addressing a widely occurring information task. We have also presented baseline performance benchmarks and introduced an information-augmented human evaluation framework that improves the reliability of QA annotations. However, manual failure analysis revealed remaining challenges in the evaluation of convincing but factually incorrect model answers.

Our research highlights the complexity of creating reliable benchmarks for non-factoid QA and the need for more sophisticated evaluation frameworks. The developed dataset and evaluation methodologies provide a strong foundation for future advancements in the field, yet further work is necessary to address the identified issues with factual consistency and overall answer coherency.

Chapter 7

Conclusions and Future Work

In recent years, significant advancements in information retrieval and natural language processing have transformed how users access information. Among these developments, question answering systems stand out for their complexity and impact, increasingly offering direct answers to user queries rather than traditional search result listings. Factoid questions, which seek specific and concise factual information (e.g., dates, names, or locations), have seen substantial progress, achieving high accuracy and reliability. In contrast, non-factoid questions present unique challenges as they demand detailed, context-rich responses that often include explanations, opinions, or multi-layered descriptions.

These complex non-factoid questions remain a significant challenge for current systems. Search result pages often lack concise or accurate answer snippets for such queries, particularly those requiring detailed or nuanced responses, forcing users to sift through multiple retrieved pages. Even when systems provide answers, verifying their correctness is difficult due to the complexity and length of the responses. While LLMs excel at generating coherent and contextually rich answers, they are prone to hallucinations, necessitating cross-checking with multiple sources. This verification process can overwhelm even expert users, as modern systems frequently produce convincing yet inaccurate answers, posing a risk of misinformation.

This thesis addresses several fundamental challenges in NFQA systems, focusing on user interaction, evaluation methodologies, and the creation of benchmarks for multi-document NFQA.

To explore the interaction behaviors with NFQA systems, we conducted an editorial

eye-tracking study. To address the first part of RQ1, focusing on how non-factoid answers are evaluated and the key influencing criteria, Chapter 3 examined how users interact with passage-level answers to non-factoid questions, analyzing the influence of answer quality, users' prior knowledge, and answer presentation. We observed that users' interest in the question and their perception of its difficulty influenced the accuracy of their ratings. Furthermore, it was more difficult for users to identify incorrect elements of an answer than to confirm its correctness. Highlighting important words in answers helps users rate answers more quickly and with less effort while maintaining quality. These results were consistent with prior studies and guided the design of experiments in subsequent research. Our study adds a novel dimension by comparing human gaze patterns with transformer attention, showing that the attention mechanism of a fine-tuned BERT model closely aligns with user gaze patterns compared to simpler TF-IDF vectors, suggesting that transformer-based models can be used to identify key information in answers effectively. Finally, we found that the completeness of an answer significantly influences its perceived correctness. This observation motivated further research into additional aspects of answer utility, leading to the development of a comprehensive taxonomy of twelve answer utility aspects in Chapter 4.

In Chapter 4, addressing the second part of RQ1 on what constitutes an optimal NFQ answer, we confirmed that relevance, correctness, completeness, and comprehensiveness are critical aspects of answer utility. Our evaluation of common answer quality metrics (ROUGE, BLEU, METEOR, and BERTScore) revealed a need for more effective metrics tailored for NFQA. Results from Chapter 4 motivated us to study variations in answer structures and evaluation processes for different categories of non-factoid questions. In Chapter 5, addressing RQ2 on how NFQs should be categorized, we developed the first streamlined taxonomy of non-factoid question categories and defined the expected answer forms for each category. This taxonomy was validated through both editorial and crowdsourcing studies, ensuring its relevance and applicability in real-world scenarios. Additionally, we introduced a new dataset of categorized non-factoid questions and a high-performance classifier to facilitate broader application and study of NFQA.

Our analysis in Chapter 5, in accordance with the second part of RQ2 on the suitability of current NFQA benchmarks, highlighted the skewed distribution of question categories in existing QA datasets, emphasizing the need for new datasets that cover a wider range of question types. We also evaluated the performance of state-of-the-art NFQA systems across different categories, revealing uneven performance and suggesting the necessity for more balanced benchmarks. These findings motivated our final research focus in Chapter 6, which addressed RQ3 in establishing a comprehensive benchmark for MD-NFQA ensuring quality, balance, and factual grounding. We developed a new dataset specifically targeting

the challenging subtask of multi-document NFQA within the instruction category of NFQA. This benchmark, WikiHowQA, is a comprehensive resource for multi-document non-factoid QA derived from WikiHow, featuring passage-level answers to instructional questions. This dataset addresses the unique challenges of multi-document question answering by providing human-written answers and supporting documents. Alongside the dataset, we conducted an analysis of human annotator performance to establish baseline benchmarks. This chapter also addresses RQ4 by assessing human capability in annotating the quality of NFQA. We proposed an information-augmented human evaluation framework to improve the reliability of long-form QA annotations, which provides valuable insights into the human evaluation process and identifies potential areas for improvement.

The findings from this thesis collectively contribute to advancing the field of non-factoid question answering by enhancing our understanding of user interaction, developing comprehensive evaluation methodologies, and creating valuable resources for further research. These contributions pave the way for future work aimed at refining evaluation frameworks, improving system performance, and expanding the scope of NFQA to address the complexities of real-world information needs.

The findings from this thesis collectively advance the field of non-factoid question answering by enhancing our understanding of user interaction, developing comprehensive evaluation methodologies, and creating valuable resources for further research. These contributions lay a solid foundation for refining evaluation frameworks—becoming essential in the era of LLM-generated answers—helping users identify correct answers. Additionally, these insights contribute to improving system performance and expanding the scope of NFQA to address the complexities of real-world information needs.

The following sections will outline the implications of these findings and propose directions for future work in NFQA and MD-NFQA.

7.1 Summary of Contributions

In this section we provide a summary of the thesis contributions by chapter.

Chapter 1 — Introduction: This chapter outlined our motivations for the thesis, defined the research scope, and provided an overview of the challenges in NFQA and MD-NFQA, as well as the thesis contributions.

Chapter 2 — Background: We reviewed the background relevant to this thesis, including the evolution of QA systems, existing taxonomies of question types, available NFQA benchmarks, and datasets for the MD-NFQA subtask. The chapter also discussed NFQA evaluation methods and user interaction with NFQA.

Chapter 3 — Exploring NFQA Evaluation Through Eye-Tracking: Human vs.

Neural Network Attention:

- Identified how users evaluate answer quality based on factors such as completeness, correctness, and presentation.
- Compared human gaze patterns with transformer attention, showing that fine-tuned BERT models align closely with user gaze patterns.
- Demonstrated that highlighting important words in NFQ answers helps users evaluate answers more efficiently and made a new highlighting approach based on BERT model weights publicly available.

Chapter 4 — Aspects of Human-Perceived Answer Utility in NFQA:

- Developed and validated a taxonomy capturing various aspects of answer utility, confirming the importance of relevance, correctness, completeness, and comprehensiveness.
- Developed and open-sourced a human-labeled NFQ answer aspects dataset.
- Assessed the effectiveness of common answer quality metrics (ROUGE, BLEU, METEOR, BERTScore), revealing the need for more tailored metrics for NFQA.

Chapter 5 — A Non-Factoid Question-Answering Taxonomy:

- Developed and validated the first streamlined taxonomy of non-factoid question categories and answer forms.
- Introduced a new dataset of categorized non-factoid questions and a high-performance classifier to aid broader research and application.
- Discovered the skewed distribution of question categories in existing QA datasets and unbalanced performance of state-of-the-art systems, emphasizing the need for more category-balanced datasets.

Chapter 6 — A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering:

- Introduced a comprehensive dataset for MD-NFQA derived from WikiHow, featuring human-written passage-level answers to instructional questions and established baseline performance benchmarks for the dataset.
- Proposed an information-augmented human evaluation framework to improve the reliability of long-form QA annotations.

7.2 Future Research

The findings from this thesis have opened several avenues for future research in NFQA. A key focus should be enhancing human evaluation frameworks, both for general long-form QA and specifically for NFQA challenges. Reliable and robust human evaluation methods remain crucial as they set the gold standard for assessment and provide a foundation for refining automatic metrics. Developing task-specific evaluation methodologies is essential to address the nuances of NFQA scenarios. For instance, evaluating instructional questions should consider the logical sequence and feasibility of steps, while assessing answers to debatable questions should prioritize balance and inclusion of diverse viewpoints to mitigate bias. Similarly, comparative questions must ensure comprehensive coverage of all aspects being compared. Such methodologies are vital for systematically and thoroughly addressing all critical aspects of the evaluated responses.

Creating frameworks that adjust evaluation criteria based on individual user preferences and needs can enhance the personalization and accuracy of evaluations. Integrating real-time fact-checking mechanisms into QA systems to verify the accuracy of generated answers, and developing models specifically designed to detect and mitigate hallucinations in generated answers, are also important steps. These models should not only flag potentially inaccurate information but also provide explanations for their assessments, enhancing the reliability of QA systems and the ability of users to verify answers.

Enhancing transparency in QA systems by providing users with information about the sources and reliability of the answers, and educating users on how to critically assess the information they receive, can help mitigate the impact of hallucinations. Additionally, incorporating multimedia elements such as images, videos, and interactive graphs will enrich the presentation of non-factoid answers further, making it important to establish reliable multi-media non-factoid answer presentations. Developing techniques that adapt the length and detail of answers based on user preferences and the complexity of the question can further improve the user experience.

Once a robust human evaluation system is in place, the development of automatic metrics that effectively capture human judgments can be pursued. In this direction, evaluating LLMs using other LLMs has been evolving as a promising approach, gaining increased traction in recent years [57, 153]. Its applicability certainly should be researched in the NFQA area.

Critically, while these methods show promise, they raise important questions about reliability, fairness, and alignment with human evaluative criteria [5]. Adapting such LLM-based frameworks to NFQA and ensuring they align closely with robust human evaluation and outperform baseline metrics is a significant step toward validating and developing

superior NFQA models. Additionally, hybrid approaches that combine human and AI judgments can enhance relevance assessments, as highlighted by recent perspectives on collaborative relevance judgments [58].

Another promising direction in automatic evaluation involves developing context-aware and user-centric metrics. Metrics that evaluate the context and coherence of answers in NFQA and incorporate real-time user interaction data can make these evaluations more responsive and accurate. For example, if overall debatable questions require balanced opinions, the answers could be adjusted to the preferences of specific users in some contexts. Additionally, creating automatic metrics tailored to specific NFQA tasks, such as instructional content or debates, will address the unique requirements and challenges of each task type, ensuring precise and meaningful assessments.

Developing algorithms that better synthesize information from multiple documents, focusing on consistency and factual accuracy, is also a promising research area. This includes improving the ability of QA systems to perform reasoning across multiple documents, identifying and resolving conflicts between different sources, and providing cohesive and accurate answers in accordance with newly established evaluation criteria. Ensuring that the context from multiple documents is preserved in the final answer is another important aspect. Developing interactive evaluation tools that allow users to provide feedback on answers can refine and improve the evaluation framework continuously.

As demonstrated in this thesis, building balanced datasets that include a wider variety of question types and domains is essential. This could involve partnerships with different organizations to gather data from diverse sources, ensuring a balanced representation of questions. Conducting studies to track the evolution of user queries and answers over time can provide insights into newer trends, ensuring that datasets remain relevant and comprehensive while avoiding the risk of models being tested on potentially seen data during pre-training.

By addressing these areas, future work can build on the contributions of this thesis to enhance the capabilities and reliability of non-factoid question answering systems, ultimately improving the user experience and finally achieving reliable, verifiable, and trustworthy question answering systems capable of handling almost any complex question a specific user might need.

Bibliography

- [1] Heba Abdel-Nabi, Arafat Awajan, and Mostafa Z. Ali. Deep learning-based question answering: a survey. *Knowl. Inf. Syst.*, 65(4):1399–1485, 2023. doi: 10.1007/S10115-022-01783-5. URL <https://doi.org/10.1007/s10115-022-01783-5>.
- [2] Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning to find answers to questions on the web. *ACM Trans. Internet Techn.*, 4(2):129–162, 2004. doi: 10.1145/990301.990303. URL <https://doi.org/10.1145/990301.990303>.
- [3] Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennett. Search, interrupted: understanding and predicting search task continuation. In William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 315–324. ACM, 2012. doi: 10.1145/2348283.2348328. URL <https://doi.org/10.1145/2348283.2348328>.
- [4] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evinaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2623–2631. ACM, 2019. doi: 10.1145/3292500.3330701. URL <https://doi.org/10.1145/3292500.3330701>.
- [5] Omar Alonso, Preetam Dammu, and Diji Yang. An interpretable an-

- swer scoring framework. In *SIGIR 2024 Workshop on Generative Information Retrieval*, 2024. URL <https://www.amazon.science/publications/an-interpretable-answer-scoring-framework>.
- [6] Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Conference on Empirical Methods in Natural Language Processing*, 2018.
- [7] Talita Anthonio, Irshad Bhat, and Michael Roth. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.702>.
- [8] Ioannis Arapakis, Filipa Peleja, Barla Berkant, and Joao Magalhaes. Linguistic benchmarks of online news article quality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1893–1902, Berlin, Germany, 2016. Association for Computational Linguistics.
- [9] Javier Artiles, Enrique Amigó, and Julio Gonzalo. The role of named entities in web people search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 534–542. ACL, 2009. URL <https://aclanthology.org/D09-1056/>.
- [10] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.
- [11] Jane Ashby, Keith Rayner, and Charles Clifton. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology*, 58(6):1065–1086, 2005.
- [12] Lorena Leal Bando, Falk Scholer, and Andrew Turpin. Constructing query-biased summaries: a comparison of human and system generated snippets. In Nicholas J. Belkin and Diane Kelly, editors, *Information Interaction in Context Symposium, IIiX 2010, New Brunswick, NJ, USA, August 18-21, 2010*, pages 195–204. ACM, 2010. doi: 10.1145/1840784.1840813. URL <https://doi.org/10.1145/1840784.1840813>.
- [13] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.
- [14] Catarina G. Belem, Pouya Pezeskhpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Rruschka. From single to multi: How llms hallucinate in multi-document summarization. *CoRR*, abs/2410.13961, 2024. doi: 10.48550/ARXIV.2410.13961. URL <https://doi.org/10.48550/arXiv.2410.13961>.
 - [15] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL <https://arxiv.org/abs/2004.05150>.
 - [16] Irshad Bhat, Talita Anthonio, and Michael Roth. Towards modeling revision requirements in wikihow instructions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8407–8414. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.675. URL <https://doi.org/10.18653/v1/2020.emnlp-main.675>.
 - [17] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. SubjQA: A dataset for subjectivity and review comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.442. URL <https://www.aclweb.org/anthology/2020.emnlp-main.442>.
 - [18] Valeria Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. In Mathieu d’Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 85–94. ACM, 2020. doi: 10.1145/3340531.3412043. URL <https://doi.org/10.1145/3340531.3412043>.
 - [19] Rishi Bommasani and Claire Cardie. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online, November 2020. Association

- for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.649. URL <https://aclanthology.org/2020.emnlp-main.649>.
- [20] Odellia Boni, Guy Feigenblat, Guy Lev, Michal Shmueli-Scheuer, Benjamin Sznaajder, and David Konopnicki. Howsumm: A multi-document summarization dataset derived from wikihow articles. *CoRR*, abs/2110.03179, 2021. URL <https://arxiv.org/abs/2110.03179>.
- [21] Pia Borlund. The concept of relevance in IR. *J. Assoc. Inf. Sci. Technol.*, 54(10): 913–925, 2003.
- [22] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki. Question answering systems: survey and trends. *Procedia Computer Science*, 73:366–375, 2015. doi: 10.1016/j.procs.2015.12.005.
- [23] Adrian M. P. Brasoveanu and Razvan Andonie. Visualizing transformers for NLP: A brief survey. In Ebad Banissi, Farzad Khosrow-shahi, Anna Ursyn, Mark W. McK. Bannatyne, João Moura Pires, Nuno Datia, Kawa Nazemi, Boris Kovalerchuk, John Counsell, Andrew Agapiou, Zora Vrcelj, Hing-Wah Chau, Mengbi Li, Gehan Nagy, Richard Laing, Rita Francese, Muhammad Sarfraz, Fatma Bouali, Gilles Venturini, Marjan Trutschl, Urska Cvek, Heimo Müller, Minoru Nakayama, Marco Temperini, Tania Di Mascio, Filippo Sciarrone, Veronica Rossano, Ralf Dörner, Loredana Caruccio, Autilia Vitiello, Weidong Huang, Michele Risi, Ugo Erra, Razvan Andonie, Muhammad Aurangzeb Ahmad, Ana Figueiras, Alfredo Cuzocrea, and Mabule Samuel Mabakane, editors, *24th International Conference on Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*, pages 270–279. IEEE, 2020. doi: 10.1109/IV51561.2020.00051. URL <https://doi.org/10.1109/IV51561.2020.00051>.
- [24] Manvi Breja and Sanjay Jain. A survey on non-factoid question answering systems. *International Journal of Computers and Applications*, 44:1–8, 07 2021. doi: 10.1080/1206212X.2021.1949117.
- [25] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002. ISSN 0163-5840. doi: 10.1145/792550.792552. URL <https://doi-org.ezproxy.lib.rmit.edu.au/10.1145/792550.792552>.
- [26] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon

- Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>.
- [27] Fan Bu, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu. Function-based question classification for general QA. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1128, Cambridge, MA, October 2010. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D10-1109>.
- [28] John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strazalkowski, Ellen Voorhees, and Ralph Weischedel. Issues, tasks and program structures to roadmap research in question answering (qa). In *Document Understanding Conference*. NIST, NIST, October 2003. URL <https://www.microsoft.com/en-us/research/publication/issues-tasks-and-program-structures-to-roadmap-research-in-question-answering/>
- [29] Christopher S. Campbell and Paul P. Maglio. A robust algorithm for reading detection. In *Proceedings of the 2001 workshop on Perceptive user interfaces, PUI '01, Orlando, Florida, USA, November 15-16, 2001*, pages 3:1–3:7. ACM, 2001. doi: 10.1145/971478.971503. URL <https://doi.org/10.1145/971478.971503>.
- [30] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer, 2013. doi: 10.1007/978-3-642-37456-2_14. URL https://doi.org/10.1007/978-3-642-37456-2_14.
- [31] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation:

- A survey. *CoRR*, abs/2006.14799, 2020. URL <https://arxiv.org/abs/2006.14799>.
- [32] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45, 2024. doi: 10.1145/3641289. URL <https://doi.org/10.1145/3641289>.
- [33] Snigdha Chaturvedi, Vittorio Castelli, Radu Florian, Ramesh M. Nallapati, and Hema Raghavan. Joint question clustering and relevance prediction for open domain non-factoid question answering. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW ’14, page 503–514, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2567999. URL <https://doi-org.ezproxy.lib.rmit.edu.au/10.1145/2566486.2567999>.
- [34] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 782–791. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00084. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html.
- [35] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Evaluating question answering evaluation. In Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen, editors, *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 119–124. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-5817. URL <https://doi.org/10.18653/v1/D19-5817>.
- [36] Long Chen, Dell Zhang, and Levene Mark. Understanding user intent in community question answering. In *Proceedings of the 21st International Conference on World Wide Web*, WWW ’12 Companion, page 823–828, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312301. doi: 10.1145/2187980.2188206. URL <https://doi-org.ezproxy.lib.rmit.edu.au/10.1145/2187980.2188206>.
- [37] Lydia B. Chilton and Jaime Teevan. Addressing people’s information needs directly in a web search result page. In Sadagopan Srinivasan, Krithi Ramamritham, Arun

- Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 27–36. ACM, 2011. doi: 10.1145/1963405.1963413. URL <https://doi.org/10.1145/1963405.1963413>.
- [38] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1300. URL <https://doi.org/10.18653/v1/n19-1300>.
- [39] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7282–7296. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.565. URL <https://doi.org/10.18653/v1/2021.acl-long.565>.
- [40] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics, 2019. doi: 10.18653/V1/W19-4828. URL <https://doi.org/10.18653/v1/W19-4828>.
- [41] Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 358–365. ACM, 2001. doi: 10.1145/383952.384024. URL <https://doi.org/10.1145/383952.384024>.

- [42] Charles L. A. Clarke, Eugene Agichtein, Susan T. Dumais, and Ryen W. White. The influence of caption features on clickthrough patterns in web search. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 135–142. ACM, 2007. doi: 10.1145/1277741.1277767. URL <https://doi.org/10.1145/1277741.1277767>.
- [43] Daniel Cohen and W. Bruce Croft. End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR ’16, page 143–146, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344975. doi: 10.1145/2970398.2970438. URL <https://doi-org.ezproxy.lib.rmit.edu.au/10.1145/2970398.2970438>.
- [44] Daniel Cohen, Liu Yang, and W. Bruce Croft. Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1165–1168. ACM, 2018. doi: 10.1145/3209978.3210118. URL <https://doi.org/10.1145/3209978.3210118>.
- [45] Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 212–219, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.28. URL <https://aclanthology.org/2021.acl-short.28>.
- [46] Eduardo Gabriel Cortes, Vinicius Woloszyn, Dante Barone, Sebastian Möller, and Renata Vieira. A systematic review of question answering systems for non-factoid questions. *J. Intell. Inf. Syst.*, 58(3):453–480, 2022. doi: 10.1007/S10844-021-00655-8. URL <https://doi.org/10.1007/s10844-021-00655-8>.
- [47] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In Mary Beth Rosson and David J. Gilmore, editors, *Proceedings of the 2007 Conference on Human Factors in Computing*

- Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 407–416. ACM, 2007. doi: 10.1145/1240624.1240690. URL <https://doi.org/10.1145/1240624.1240690>.
- [48] Hoa Trang Dang. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, Sydney, Australia, July 2006. Association for Computational Linguistics. URL <https://aclanthology.org/W06-0707>.
- [49] Daniel Deutsch and Dan Roth. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.24. URL <https://aclanthology.org/2021.conll-1.24>.
- [50] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Trans. Assoc. Comput. Linguistics*, 9:774–789, 2021. doi: 10.1162/tacl_a_00397. URL https://doi.org/10.1162/tacl_a_00397.
- [51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL <https://doi.org/10.18653/v1/n19-1423>.
- [52] Eleftherios Dimitrakis, Konstantinos Sgontzos, and Yannis Tzitzikas. A survey on question answering systems over linked data and documents. *J. Intell. Inf. Syst.*, 55(2):233–259, 2020. doi: 10.1007/S10844-019-00584-7. URL <https://doi.org/10.1007/s10844-019-00584-7>.
- [53] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7250–7274. Association for Computational Linguistics,

2022. doi: 10.18653/v1/2022.acl-long.501. URL <https://doi.org/10.18653/v1/2022.acl-long.501>.
- [54] Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/L18-1438>.
- [55] Oren Etzioni. Search needs a shake-up. *Nat.*, 476(7358):25–26, 2011. doi: 10.1038/476025A. URL <https://doi.org/10.1038/476025a>.
- [56] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1102. URL <https://aclanthology.org/P19-1102>.
- [57] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi, editors, *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, pages 39–50. ACM, 2023. doi: 10.1145/3578337.3605136. URL <https://doi.org/10.1145/3578337.3605136>.
- [58] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Who determines what is relevant? humans or ai? why not both? *Commun. ACM*, 67(4):31–34, 2024. doi: 10.1145/3624730. URL <https://doi.org/10.1145/3624730>.
- [59] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL <https://www.aclweb.org/anthology/P19-1346>.

- [60] Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Laura Monceaux, Isabelle Robba, and Anne Vilnat. Finding an answer based on the recognition of the question focus. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001*, volume 500-250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2001. URL <http://trec.nist.gov/pubs/trec10/papers/qaLIR.pdf>.
- [61] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. Building watson: An overview of the deepqa project. *AI Mag.*, 31(3):59–79, 2010. doi: 10.1609/aimag.v31i3.2303. URL <https://doi.org/10.1609/aimag.v31i3.2303>.
- [62] Pnina Fichman. A comparative assessment of answer quality on four question answering sites. *J. Inf. Sci.*, 37(5):476–486, 2011.
- [63] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
- [64] Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. Higher-order lexical semantic models for non-factoid answer reranking. *Trans. Assoc. Comput. Linguistics*, 3:197–210, 2015. doi: 10.1162/TACL_A_00133. URL https://doi.org/10.1162/tacl_a_00133.
- [65] Hengyi Fu, Shuheng Wu, and Sanghee Oh. Evaluating answer quality across knowledge domains: Using textual and non-textual features in social q&a. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, USA, 2015. American Society for Information Science.
- [66] Kavita Ganeshan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 340–348. Tsinghua University Press, 2010. URL <https://aclanthology.org/C10-1039/>.
- [67] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018. URL <http://arxiv.org/abs/1803.07640>.

- [68] Jorge Martínez Gil. A survey on legal question-answering systems. *Comput. Sci. Rev.*, 48:100552, 2023. doi: 10.1016/J.COSREV.2023.100552. URL <https://doi.org/10.1016/j.cosrev.2023.100552>.
- [69] Maharshi Gor, Hal Daumé Iii, Tianyi Zhou, and Jordan Lee Boyd-Graber. Do great minds think alike? investigating human-AI complementarity in question answering with CAIMIRA. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21533–21564, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1201. URL <https://aclanthology.org/2024.emnlp-main.1201>.
- [70] Arthur C. Graesser and Natalie K. Person. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137, 1994. doi: 10.3102/00028312031001104. URL <https://doi.org/10.3102/00028312031001104>.
- [71] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in WWW search. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 478–479. ACM, 2004. doi: 10.1145/1008992.1009079. URL <https://doi.org/10.1145/1008992.1009079>.
- [72] B. F. Green, A. K. Wolf, C. Chomsky, and K. R. Laughery. Baseball. *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference on - IRE-AIEE-ACM '61 (Western)*, 1961. doi: 10.1145/1460690.1460714.
- [73] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL <https://aclanthology.org/N18-1065>.
- [74] Deepak Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Jain, and Shubhashis Sengupta. Can taxonomy help? improving semantic question matching using question taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 499–513, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL <https://aclanthology.org/C18-1042>.

- [75] Ido Guy, Victor Makarenkov, Niva Hazon, Lior Rokach, and Bracha Shapira. Identifying informational vs. conversational questions on community question answering archives. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 216–224, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159733. URL <https://doi-org.ezproxy.lib.rmit.edu.au/10.1145/3159652.3159733>.
- [76] Jacek Gwizdka. Characterizing relevance with eye-tracking measures. In David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson, editors, *Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26-29, 2014*, pages 58–67. ACM, 2014. doi: 10.1145/2637002.2637011. URL <https://doi.org/10.1145/2637002.2637011>.
- [77] Sanda M. Harabagiu, Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. FALCON: boosting knowledge for answer engines. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*, volume 500-249 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2000. URL <http://trec.nist.gov/pubs/trec9/papers/sm1.pdf>.
- [78] Kerry Hart and Anita Sarma. Perceptions of answer quality in an online technical question and answer forum. In *Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering*, pages 103–106, New York, NY, USA, 2014. Association for Computing Machinery.
- [79] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. Performance prediction for non-factoid question answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 55–58, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368810. doi: 10.1145/3341981.3344249. URL <https://doi.org/10.1145/3341981.3344249>.
- [80] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. Performance prediction for non-factoid question answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 55–58, New York, NY, USA, 2019. ACM.
- [81] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang.

- Dureader: a chinese machine reading comprehension dataset from real-world applications. In Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant, editors, *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 37–46. Association for Computational Linguistics, 2018. doi: 10.18653/V1/W18-2605. URL <https://aclanthology.org/W18-2605/>.
- [82] Suzanne Hidi and K. Ann Renninger. The four-phase model of interest development. *Educational Psychologist*, 41(2):111–127, 2006. doi: 10.1207/s15326985ep4102_4.
- [83] Ryuichiro Higashinaka and Hideki Isozaki. Corpus-based question answering for why-questions. In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 418–425. The Association for Computer Linguistics, 2008. URL <https://aclanthology.org/I08-1055/>.
- [84] Lynette Hirschman and Robert J. Gaizauskas. Natural language question answering: the view from here. *Nat. Lang. Eng.*, 7(4):275–300, 2001. doi: 10.1017/S1351324901002807. URL <https://doi.org/10.1017/S1351324901002807>.
- [85] Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920, 2017. doi: 10.3233/SW-160247. URL <https://doi.org/10.3233/SW-160247>.
- [86] Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Found. Trends Inf. Retr.*, 10(1):1–117, 2016. doi: 10.1561/1500000051. URL <https://doi.org/10.1561/1500000051>.
- [87] Eduard Hovy, Ul Hermjakob, and Deep Ravichandran. A question/answer typology with surface text patterns. 01 2002. doi: 10.3115/1289189.1289206.
- [88] David A. Hull. Xerox TREC-8 question answering track report. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 1999. URL <http://trec.nist.gov/pubs/trec8/papers/xerox-QA.pdf>.
- [89] Tereza Iofciu, Nick Craswell, and Milad Shokouhi. Evaluating the impact of snippet highlighting in search. *UIIR-2009*, page 44, 2009.

- [90] Mohit Iyyer, Jordan L. Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 633–644. ACL, 2014. doi: 10.3115/V1/D14-1070. URL <https://doi.org/10.3115/v1/d14-1070>.
- [91] Peter Jansen, Mihai Surdeanu, and Peter Clark. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 977–986. The Association for Computer Linguistics, 2014. doi: 10.3115/V1/P14-1092. URL <https://doi.org/10.3115/v1/p14-1092>.
- [92] Zhiling Jin, Yu Hong, Hongyu Zhu, Jianmin Yao, and Min Zhang. Bi-granularity adversarial training for non-factoid answer retrieval. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 322–335. Springer, 2022. doi: 10.1007/978-3-030-99736-6_22. URL https://doi.org/10.1007/978-3-030-99736-6_22.
- [93] Thorsten Joachims and Filip Radlinski. Search engines that learn from implicit feedback. *Computer*, 40(8):34–40, 2007.
- [94] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL <https://doi.org/10.18653/v1/P17-1147>.
- [95] Lakshmi K and Sindhu S. A study on various approaches towards non-factoid question answering systems. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 780–788, 2023. doi: 10.1109/ICAAIC56838.2023.10140381.

- [96] Johanna K Kaakinen and Jukka Hyönä. Perspective effects in repeated reading: An eye movement study. *Memory & Cognition*, 35(6):1323–1336, 2007.
- [97] Ehsan Kamalloo, Charles L. A. Clarke, and Davood Rafiei. Limitations of open-domain question answering benchmarks for document-level reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2123–2128, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3592011. URL <https://doi.org/10.1145/3539618.3592011>.
- [98] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.307. URL <https://aclanthology.org/2023.acl-long.307>.
- [99] Chris Kamphuis, Aileen Lin, Siwen Yang, Jimmy Lin, Arjen P. de Vries, and Faegheh Hasibi. MMEA: MS MARCO entity annotations and disambiguations. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2817–2825. ACM, 2023. doi: 10.1145/3539618.3591887. URL <https://doi.org/10.1145/3539618.3591887>.
- [100] Mostafa Keikha, Jae Hyun Park, W. Bruce Croft, and Mark Sanderson. Retrieving passages and finding answers. In J. Shane Culpepper, Laurence Anthony F. Park, and Guido Zuccon, editors, *Proceedings of the 2014 Australasian Document Computing Symposium, ADCS 2014, Melbourne, VIC, Australia, November 27-28, 2014*, page 81. ACM, 2014. doi: 10.1145/2682862.2682877. URL <https://doi.org/10.1145/2682862.2682877>.
- [101] Diane Kelly and Colleen Cool. The effects of topic familiarity on information search behavior. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2002, Portland, Oregon, USA, June 14-18, 2002, Proceedings*, pages 74–75. ACM, 2002. doi: 10.1145/544220.544232. URL <https://doi.org/10.1145/544220.544232>.
- [102] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi

- Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4110–4124. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.324. URL <https://doi.org/10.18653/v1/2021.nacl-main.324>.
- [103] Soojung Kim and Sanghee Oh. Users' relevance criteria for evaluating answers in a social q&a site. *J. Assoc. Inf. Sci. Technol.*, 60(4):716–727, 2009. doi: 10.1002/ASI.21026. URL <https://doi.org/10.1002/asi.21026>.
- [104] Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Inf. Sci.*, 181(24):5412–5434, 2011. doi: 10.1016/J.INS.2011.07.047. URL <https://doi.org/10.1016/j.ins.2011.07.047>.
- [105] Mahnaz Koupaei and William Yang Wang. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305, 2018. URL <http://arxiv.org/abs/1810.09305>.
- [106] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4940–4957. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.nacl-main.393. URL <https://doi.org/10.18653/v1/2021.nacl-main.393>.
- [107] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1051. URL <https://aclanthology.org/D19-1051>.

- [108] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.750. URL <https://doi.org/10.18653/v1/2020.emnlp-main.750>.
- [109] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatsmanoglu. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 633–642, New York, NY, USA, 2012. Association for Computing Machinery.
- [110] Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *CoRR*, abs/2010.12694, 2020. URL <https://arxiv.org/abs/2010.12694>.
- [111] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL <https://aclanthology.org/Q19-1026>.
- [112] Cody C. T. Kwok, Oren Etzioni, and Daniel S. Weld. Scaling question answering to the web. In Vincent Y. Shen, Nobuo Saito, Michael R. Lyu, and Mary Ellen Zurko, editors, *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, pages 150–161. ACM, 2001. doi: 10.1145/371920.371973. URL <https://doi.org/10.1145/371920.371973>.
- [113] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.360. URL <https://aclanthology.org/2020.findings-emnlp.360>.
- [114] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In Shlomo Geva,

- Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 113–122. ACM, 2014. doi: 10.1145/2600428.2609631. URL <https://doi.org/10.1145/2600428.2609631>.
- [115] Weronika Lajewska and Krisztian Balog. Towards reliable and factual response generation: Detecting unanswerable questions in information-seeking conversations. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part III*, volume 14610 of *Lecture Notes in Computer Science*, pages 336–344. Springer, 2024. doi: 10.1007/978-3-031-56063-7_25. URL https://doi.org/10.1007/978-3-031-56063-7_25.
- [116] Long T. Le, Chirag Shah, and Erik Choi. Evaluating the quality of educational answers in community question-answering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 129–138, New York, NY, USA, 2016. ACM.
- [117] Guang-He Lee and Yun-Nung Chen. MUSE: Modularizing unsupervised sense embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1034. URL <https://www.aclweb.org/anthology/D17-1034>.
- [118] Shun-Yang Lee, Huaxia Rui, and Andrew B. Whinston. Is best answer really the best answer? the politeness bias. *MIS Q.*, 43(2):579–600, 2019.
- [119] Wendy G. Lehnert. A conceptual theory of question answering. In Raj Reddy, editor, *Proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, USA, August 22-25, 1977*, pages 158–164. William Kaufmann, 1977.
- [120] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://www.aclweb.org/anthology/2020.acl-main.703>.

- [121] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [122] Patrick S. H. Lewis, Pontus Stenetorp, and Sebastian Riedel. Question and answer test-train overlap in open-domain question answering datasets. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1000–1008. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.86. URL <https://doi.org/10.18653/v1/2021.eacl-main.86>.
- [123] Xin Li and Dan Roth. Learning question classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*, 2002. URL <https://aclanthology.org/C02-1150/>.
- [124] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.
- [125] Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. What makes a good answer? the role of context in question answering. In Matthias Rauterberg, Marino Menozzi, and Janet Wesson, editors, *Human-Computer Interaction INTERACT ’03: IFIP TC13 International Conference on Human-Computer Interaction, 1st-5th September 2003, Zurich, Switzerland*. IOS Press, 2003.
- [126] Kenneth C. Litkowski. Summarization experiments in duc 2004. 2004.
- [127] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.

- [128] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.
- [129] Yao Lu, Yue Dong, and Laurent Charlin. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.648. URL <https://aclanthology.org/2020.emnlp-main.648>.
- [130] Yosi Mass, Haggai Roitman, Shai Erera, Or Rivlin, Bar Weiner, and David Konopnicki. A study of bert for non-factoid question-answering under passage length constraints. *ArXiv*, abs/1908.06780, 2019.
- [131] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html>.
- [132] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748. URL <https://doi.org/10.1145/219717.219748>.
- [133] Amit Mishra and Sanjay Kumar Jain. A survey on question answering systems with classification. *J. King Saud Univ. Comput. Inf. Sci.*, 28(3):345–361, 2016. doi: 10.1016/J.JKSUCI.2014.10.007. URL <https://doi.org/10.1016/j.jksuci.2014.10.007>.
- [134] Diganta Misra. Mish: A self regularized non-monotonic activation function. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. URL <https://www.bmvc2020-conference.com/assets/papers/0928.pdf>.
- [135] Junta Mizuno, Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. Non-factoid question answering experiments at NTCIR-6: towards answer type detection for realworld questions. In Noriko Kando, editor, *Proceedings of the 6th NTCIR Workshop Meeting*

- on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-6, National Center of Sciences, Tokyo, Japan, May 15-18, 2007.* National Institute of Informatics (NII), 2007. URL <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/71.pdf>.
- [136] Jerome L. Myers, Arnold D. Well, and Robert F. Lorch. *Research Design and Statistical Analysis*. Routledge, 2010.
- [137] Leann Myers and Maria J. Sirois. Differences between spearman correlation coefficients. In *Encyclopedia of Statistical Sciences*. John Wiley and Sons, 2006.
- [138] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021. URL <https://arxiv.org/abs/2112.09332>.
- [139] Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1098. URL <https://aclanthology.org/P17-1098>.
- [140] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d’Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.
- [141] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine reading COmprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d’Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1773. CEUR-WS.org, 2016.

- [142] Vincent Nguyen. Question answering in the biomedical domain. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 54–63. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-2008. URL <https://doi.org/10.18653/v1/p19-2008>.
- [143] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. A semi-supervised learning approach to why-question answering. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3022–3029. AAAI Press, 2016. doi: 10.1609/AAAI.V30I1.10388. URL <https://doi.org/10.1609/aaai.v30i1.10388>.
- [144] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, USA, 2002. Association for Computational Linguistics.
- [145] Jae Hyun Park and W. Bruce Croft. Using key concepts in a translation model for retrieval. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto, editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 927–930. ACM, 2015. doi: 10.1145/2766462.2767768. URL <https://doi.org/10.1145/2766462.2767768>.
- [146] Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. Data augmentation for abstractive query-focused multi-document summarization. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13666–13674. AAAI Press, 2021. doi: 10.1609/AAAI.V35I15.17611. URL <https://doi.org/10.1609/aaai.v35i15.17611>.
- [147] Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. Data augmentation for abstractive query-focused multi-document summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13666–13674, 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17611>.

- [148] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. 46(4):1023–1031, 2014. ISSN 1554-3528. doi: 10.3758/s13428-013-0434-y. URL <https://doi.org/10.3758/s13428-013-0434-y>.
- [149] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2523–2544. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.nacl-main.200. URL <https://doi.org/10.18653/v1/2021.nacl-main.200>.
- [150] John M. Prager, Eric W. Brown, Anni Coden, and Dragomir R. Radev. Question-answering by predictive annotation. In Emmanuel J. Yannakoudakis, Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, pages 184–191. ACM, 2000. doi: 10.1145/345508.345574. URL <https://doi.org/10.1145/345508.345574>.
- [151] Susan Prion and Katie Haerling. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10:535–536, 2014.
- [152] Chen Qu, Liu Yang, W. Bruce Croft, Falk Scholer, and Yongfeng Zhang. Answer interaction in non-factoid question answering systems. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 249–253, New York, NY, USA, 2019. ACM.
- [153] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. Llm4eval: Large language model for evaluation in ir. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 3040–3043, New York, NY, USA, 2024.

- Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657992. URL <https://doi.org/10.1145/3626772.3657992>.
- [154] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://www.aclweb.org/anthology/D16-1264>.
- [155] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics.
- [156] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL <https://www.aclweb.org/anthology/P18-2124>.
- [157] Keith Rayner, Erik D Reichle, Michael J Stroud, Carrick C Williams, and Alexander Pollatsek. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and aging*, 21(3):448, 2006.
- [158] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW ’04, page 13–19, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 158113844X. doi: 10.1145/988672.988675. URL <https://doi.org/10.1145/988672.988675>.
- [159] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D07-1043>.
- [160] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:103–117, 2021. doi: 10.1162/tacl_a_00400. URL https://doi.org/10.1162/tacl_a_00400.

- citation for Computational Linguistics, 9:53–68, 2021. doi: 10.1162/tacl_a_00353.
URL <https://aclanthology.org/2021.tacl-1.4>.
- [161] Andreas Rücklé and Iryna Gurevych. Representation learning for answer selection with lstm-based importance weighting. In Claire Gardent and Christian Retoré, editors, *IWCS 2017 - 12th International Conference on Computational Semantics - Short papers, Montpellier, France, September 19 - 22, 2017*. The Association for Computer Linguistics, 2017. URL <https://aclanthology.org/W17-6935/>.
- [162] Andreas Rücklé and Iryna Gurevych. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In Mohit Bansal and Heng Ji, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pages 19–24. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-4004. URL <https://doi.org/10.18653/v1/P17-4004>.
- [163] Tefko Saracevic. *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?* Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2016.
- [164] Denis Savenkov, Pavel Braslavski, and Mikhail Lebedev. Search snippet evaluation at yandex: Lessons learned and future directions. In Pamela Forner, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas, and Maarten de Rijke, editors, *Multilingual and Multimodal Information Access Evaluation - Second International Conference of the Cross-Language Evaluation Forum, CLEF 2011, Amsterdam, The Netherlands, September 19-22, 2011. Proceedings*, volume 6941 of *Lecture Notes in Computer Science*, pages 14–25. Springer, 2011. doi: 10.1007/978-3-642-23708-9\4. URL https://doi.org/10.1007/978-3-642-23708-9_4.
- [165] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. Questeval: Summarization asks for fact-based evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.529. URL <https://doi.org/10.18653/v1/2021.emnlp-main.529>.
- [166] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL <https://aclanthology.org/P17-1099>.
- [167] Priyanka Sen and Amir Saffari. What do models learn from question answering datasets? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2429–2438. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.190. URL <https://doi.org/10.18653/v1/2020.emnlp-main.190>.
- [168] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1162. URL <https://doi.org/10.18653/v1/p16-1162>.
- [169] Chirag Shah. Building a parsimonious model for identifying best answers using interaction history in community q&a. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, USA, 2015. American Society for Information Science.
- [170] Chirag Shah and Jefferey Pomerantz. Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 411–418, New York, NY, USA, 2010. Association for Computing Machinery.
- [171] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01075. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shridhar_ALFRED_A_Benchmark_for_Interpreting_Grounded_Instructions_for_Everyday_Tasks_CVPR_2020_paper.html.
- [172] Amit Singhal, Steve Abney, Michiel Bacchiani, Michael Collins, Donald Hindle, and Fernando Pereira. Att at trec-8. 11 1999.

- [173] Amir Soleimani, Christof Monz, and Marcel Worring. NLQuAD: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.106. URL <https://aclanthology.org/2021.eacl-main.106>.
- [174] Radu Soricut and Eric Brill. Automatic question answering using the web: Beyond the factoid. *Inf. Retr.*, 9(2):191–206, 2006. doi: 10.1007/s10791-006-7149-y. URL <https://doi.org/10.1007/s10791-006-7149-y>.
- [175] K Spärck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840, 2000.
- [176] Rohini Srihari and Wei Li. A question answering system supported by information extraction. In *Sixth Applied Natural Language Processing Conference*, pages 166–172, Seattle, Washington, USA, April 2000. Association for Computational Linguistics. doi: 10.3115/974147.974170. URL <https://www.aclweb.org/anthology/A00-1023>.
- [177] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL <https://aclanthology.org/2022.emnlp-main.566>.
- [178] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online QA collections. In Kathleen R. McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 719–727. The Association for Computer Linguistics, 2008. URL <https://aclanthology.org/P08-1082/>.
- [179] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Comput. Linguistics*, 37(2):351–383, 2011. doi: 10.1162/COLI_A_00051. URL https://doi.org/10.1162/COLI_a_00051.

- [180] Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda. Question classification using HDAG kernel. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 61–68, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1119312.1119320. URL <https://www.aclweb.org/anthology/W03-1208>.
- [181] Ming Tan, Bing Xiang, and Bowen Zhou. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108, 2015. URL <http://arxiv.org/abs/1511.04108>.
- [182] Andrew Tawfik, Arthur Graesser, Jessica Gatewood, and Jaclyn Gishbaugher. Role of questions in inquiry-based instruction: towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, 68:1–25, 01 2020. doi: 10.1007/s11423-020-09738-9.
- [183] Andrew Tawfik, Arthur Graesser, Jessica Gatewood, and Jaclyn Gishbaugher. Role of questions in inquiry-based instruction: towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, pages 1–25, 2020.
- [184] Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.303. URL <https://aclanthology.org/2021.findings-acl.303>.
- [185] Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.290947. URL <https://doi.org/10.1145/290941.290947>.
- [186] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention interpretability across NLP tasks. *CoRR*, abs/1909.11218, 2019. URL <http://arxiv.org/abs/1909.11218>.
- [187] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus,

- S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fdb053c1c4a845aa-Abstract.html>.
- [188] Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. Towards argument mining for social good: A survey. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1338–1352. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.107. URL <https://doi.org/10.18653/v1/2021.acl-long.107>.
- [189] Suzan Verberne, Lou Boves, Peter-Arno Coppen, and Nelleke Oostdijk. Discourse-based answering of why-questions. *Traitemen Automatique des Langues*, 47(2):21–41, 2006. URL <http://www.atala.org/Discourse-based-answering-of-why>.
- [190] Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. Data for question answering: The case of why. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/525_pdf.pdf.
- [191] Suzan Verberne, Hans van Halteren, Daphne Theijssen, Stephan Raaijmakers, and Lou Boves. Learning to rank for *why*-question answering. *Inf. Retr.*, 14(2):107–132, 2011. doi: 10.1007/S10791-010-9136-6. URL <https://doi.org/10.1007/s10791-010-9136-6>.
- [192] Jesse Vig. A multiscale visualization of attention in the transformer model. In Marta R. Costa-jussà and Enrique Alfonseca, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 37–42. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-3007. URL <https://doi.org/10.18653/v1/p19-3007>.
- [193] MK Vijaymeena and K Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28, 2016.

- [194] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1580. URL <https://doi.org/10.18653/v1/p19-1580>.
- [195] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001*, volume 500-250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2001. URL <http://trec.nist.gov/pubs/trec10/papers/qa10.pdf>.
- [196] Jaap Walhout, Paola Oomen, Halszka Jarodzka, and Saskia Brand-Gruwel. Effects of task complexity on online search behavior of adolescents. *Journal of the Association for Information Science and Technology*, 68(6):1449–1461, 2017.
- [197] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html>.
- [198] Lu Wang and Wang Ling. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1007. URL <https://aclanthology.org/N16-1007>.
- [199] Mengqiu Wang et al. A survey of answer extraction techniques in factoid question answering. *Computational Linguistics*, 1(1):1–14, 2006.
- [200] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. Is this your final answer?: Evaluating the effect

- of answers on good abandonment in mobile search. In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 889–892. ACM, 2016. doi: 10.1145/2911451.2914736. URL <https://doi.org/10.1145/2911451.2914736>.
- [201] Gary S Wolverton and David Zola. The temporal characteristics of visual information extraction during reading. In *Eye movements in reading*, pages 41–51. 1983.
- [202] William A. Woods. Progress in natural language understanding: an application to lunar geology. In *American Federation of Information Processing Societies: 1973 National Computer Conference, 4-8 June 1973, New York, NY, USA*, volume 42 of *AFIPS Conference Proceedings*, pages 441–450. AFIPS Press/ACM, 1973. doi: 10.1145/1499586.1499695. URL <https://doi.org/10.1145/1499586.1499695>.
- [203] Zhijing Wu, Mark Sanderson, Berkant Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. Providing direct answers in search results: A study of user behavior. In Mathieu d’Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1635–1644. ACM, 2020. doi: 10.1145/3340531.3412017. URL <https://doi.org/10.1145/3340531.3412017>.
- [204] Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1496. URL <https://www.aclweb.org/anthology/P19-1496>.
- [205] Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy, 2019. Association for Computational Linguistics.
- [206] Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

- 1: Long Papers)*, pages 3556–3572, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.249. URL <https://aclanthology.org/2022.acl-long.249>.
- [207] Yunjie (Calvin) Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006.
- [208] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. Cqarank: jointly model topics and expertise in community question answering. In Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Radeev Rastogi, editors, *22nd ACM International Conference on Information and Knowledge Management, CIKM’13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 99–108. ACM, 2013. doi: 10.1145/2505515.2505720. URL <https://doi.org/10.1145/2505515.2505720>.
- [209] Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W. Bruce Croft, Jiafeng Guo, and Falk Scholer. Beyond factoid QA: effective methods for non-factoid answer sentence retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*, pages 115–128. Springer, 2016. doi: 10.1007/978-3-319-30671-1_9. URL https://doi.org/10.1007/978-3-319-30671-1_9.
- [210] Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikihow. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2167–2179. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.165. URL <https://doi.org/10.18653/v1/2021.emnlp-main.165>.
- [211] Yuan Yao, Hanghang Tong, Tao Xie, Leman Akoglu, Feng Xu, and Jian Lu. Detecting high-quality posts in community question answering sites. *Information Sciences*, 302:70–82, 2015.
- [212] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. Scisummnet: A large annotated corpus and

- content-impact models for scientific paper summarization with citation networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33017386. URL <https://doi.org/10.1609/aaai.v33i01.33017386>.
- [213] Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W. Bruce Croft, and Mark Sanderson. Document summarization for answering non-factoid queries. *IEEE Trans. Knowl. Data Eng.*, 30(1):15–28, 2018. doi: 10.1109/TKDE.2017.2754373. URL <https://doi.org/10.1109/TKDE.2017.2754373>.
- [214] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1472. URL <https://doi.org/10.18653/v1/p19-1472>.
- [215] Hui Zhang. Beyond query-oriented highlighting: Investigating the effect of snippet text highlighting in search user behavior. *Comput. Intell. Neurosci.*, 2018:7836969:1–7836969:12, 2018. doi: 10.1155/2018/7836969. URL <https://doi.org/10.1155/2018/7836969>.
- [216] Li Zhang, Qing Lyu, and Chris Callison-Burch. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.374. URL <https://aclanthology.org/2020.emnlp-main.374>.
- [217] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.
- [218] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. pages 435–444, New York, NY, USA, 2014. Association for Computing Machinery.

- [219] Mingjun Zhao, Shengli Yan, Bang Liu, Xinwang Zhong, Qian Hao, Haolan Chen, Di Niu, Bowei Long, and Weidong Guo. Qbsum: A large-scale query-based document summarization dataset from real-world applications. *Computer Speech & Language*, 66:101166, 2021. ISSN 0885-2308. doi: <https://doi.org/10.1016/j.csl.2020.101166>. URL <https://www.sciencedirect.com/science/article/pii/S0885230820300991>.
- [220] Yiming Zhao, Jin Zhang, Xue Xia, and Taowen Le. Evaluation of google question-answering quality. *Libr. Hi Tech*, 37(2):312–328, 2019. doi: 10.1108/LHT-10-2017-0218. URL <https://doi.org/10.1108/LHT-10-2017-0218>.
- [221] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. Human behavior inspired machine reading comprehension. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 425–434. ACM, 2019. doi: 10.1145/3331184.3331231. URL <https://doi.org/10.1145/3331184.3331231>.
- [222] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5905–5921. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.nacl-main.472. URL <https://doi.org/10.18653/v1/2021.nacl-main.472>.
- [223] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A multi-dimensional model for assessing the quality of answers in social q&a sites. In Paul L. Bowen, Ahmed K. Elmagarmid, Hubert Österle, and Kai-Uwe Sattler, editors, *Proceedings of the 14th International Conference on Information Quality*, pages 264–265. HPI/MIT, 2009.

Appendices

Appendix A

List of Publications

Parts of this thesis have been previously published in the following conference papers:

- Bolotova, V., Blinov, V., Zheng, Y., Croft, W. B., Scholer, F., Sanderson, M. (2020). Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid QA evaluation. Proceedings of the 29th ACM International Conference on Information Knowledge Management (CIKM '20), 85-94. <https://doi.org/10.1145/3340531.3412043>
- Cambazoglu, B.B., Bolotova-Baranova, V., Scholer, F., Sanderson, M., Tavakoli, L., Croft, W.B. (2021). Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering. In Scholer, F., Thomas, P., Elsweiler, D., Joho, H., Kando, N., Smith, C. (Eds.), CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021 (pp. 75-84). ACM. <https://doi.org/10.1145/3406522.3446028>
- Bolotova, V., Blinov, V., Scholer, F., Croft, W. B., Sanderson, M. (2022). A Non-Factoid Question-Answering Taxonomy. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1196-1207). ACM. doi: 10.1145/3477495.3531926
- Bolotova-Baranova, V., Blinov, V., Filippova, S., Scholer, F., Sanderson, M. (2023). WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering. In Proceedings of the 61st Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers) (pp. 5291-5314). Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.290

Appendix B

Ethics Approval

The following is the approval letter for the ethics application pertaining to the research conducted in Chapter 3, Chapter 4, Chapter 5, and Chapter 6 respectively.

College of Science, Health & Engineering

College Human Ethics Advisory Network (CHEAN)
Email: seh-human-ethics@rmit.edu.au
Tel: [61 3] 9925 4620
Building 91, Level 2, City Campus/Building 215, Level 2, Bundoora West Campus

Notice of Approval

Date: **16 October 2019**

Project number: **66-19/22334**

Project title: **Information storage, analysis and retrieval**

Risk classification: **Negligible risk**

Chief investigator: **Professor Mark Sanderson**

Status: **Approved**

Approval period: From: **16/10/2019** To: **16/10/2022**

The following documents have been reviewed and approved:

Title	Version	Date
Risk Assessment and Application Form	2	29 August 2019
Participant Information Sheet and Consent Form	2	29 August 2019

The above application has been approved by the RMIT University CHEAN as it meets the requirements of the *National Statement on Ethical Conduct in Human Research* (NHMRC, 2007).

Terms of approval:

1. Responsibilities of chief investigator

It is the responsibility of the above chief investigator to ensure that all other investigators and staff on a project are aware of the terms of approval and to ensure that the project is conducted as approved by CHEAN. Approval is valid only whilst the chief investigator holds a position at RMIT University.

2. Amendments

Approval must be sought from CHEAN to amend any aspect of a project. To apply for an amendment, use the request for amendment form, which is available on the HREC website and submitted to the CHEAN secretary. Amendments must not be implemented without first gaining approval from CHEAN.

3. Adverse events

You should notify the CHEAN immediately (within 24 hours) of any serious or unanticipated adverse effects of their research on participants, and unforeseen events that might affect the ethical acceptability of the project.

4. Annual reports

Continued approval of this project is dependent on the submission of an annual report. Annual reports must be submitted by the anniversary of approval of the project for each full year of the project. If the project is of less than 12 months duration, then a final report only is required.

5. Final report

A final report must be provided within six months of the end of the project. CHEAN must be notified if the project is discontinued before the expected date of completion.

6. Monitoring

Projects may be subject to an audit or any other form of monitoring by the CHEAN at any time.

7. Retention and storage of data

The investigator is responsible for the storage and retention of original data according to the requirements of the *Australian Code for the Responsible Conduct of Research* (R22) and relevant RMIT policies.

8. Special conditions of approval

Nil.

In any future correspondence please quote the project number and project title above.

Yours sincerely,

Dr Lauren Saling

Deputy Chair, Science Engineering & Health

College Human Ethics Advisory Network

College of Science, Health & Engineering

College Human Ethics Advisory Network (CHEAN)
Email: seh-human-ethics@rmit.edu.au
Tel: [61 3] 9925 4620
Building 91, Level 2, City Campus/Building 215, Level 2, Bundoora West Campus

Notice of Approval

Date: **4 October 2019**

Project number: **78-19/22427**

Project title: **Investigating the user behaviour in interactive question answer system**

Risk classification: **Low risk**

Chief investigator: **Professor Mark Sanderson**

Status: **Approved**

Approval period: From: **4/10/2019** To: **4/12/2019**

The following documents have been reviewed and approved:

Title	Version	Date
Risk Assessment and Application Form	2	4 October 2019
Participant Information Sheet and Consent Form (x2)	2	4 October 2019
Recruitment Material	1	10 September 2019
Questionnaire (x2)	1	10 September 2019

The above application has been approved by the RMIT University CHEAN as it meets the requirements of the *National Statement on Ethical Conduct in Human Research* (NHMRC, 2007).

Terms of approval:

1. Responsibilities of chief investigator

It is the responsibility of the above chief investigator to ensure that all other investigators and staff on a project are aware of the terms of approval and to ensure that the project is conducted as approved by CHEAN. Approval is valid only whilst the chief investigator holds a position at RMIT University.

2. Amendments

Approval must be sought from CHEAN to amend any aspect of a project. To apply for an amendment, use the request for amendment form, which is available on the HREC website and submitted to the CHEAN secretary. Amendments must not be implemented without first gaining approval from CHEAN.

3. Adverse events

You should notify the CHEAN immediately (within 24 hours) of any serious or unanticipated adverse effects of their research on participants, and unforeseen events that might affect the ethical acceptability of the project.

4. Annual reports

Continued approval of this project is dependent on the submission of an annual report. Annual reports must be submitted by the anniversary of approval of the project for each full year of the project. If the project is of less than 12 months duration, then a final report only is required.

5. Final report

A final report must be provided within six months of the end of the project. CHEAN must be notified if the project is discontinued before the expected date of completion.

6. Monitoring

Projects may be subject to an audit or any other form of monitoring by the CHEAN at any time.

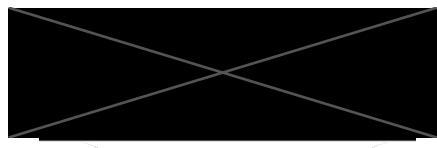
7. Retention and storage of data

The investigator is responsible for the storage and retention of original data according to the requirements of the *Australian Code for the Responsible Conduct of Research* (R22) and relevant RMIT policies.

8. Special conditions of approval

Nil.

In any future correspondence please quote the project number and project title above.



**Associate Professor Barbara Polus
Chair, Science Engineering & Health
College Human Ethics Advisory Network**

Cc Student Investigator/s: Zhijing Wu, Yuken Zheng
Other investigator/s: A/Prof Falk Scholer, Dr Berkant Cambazoglu

STEM College

College Human Ethics Advisory
Network (CHEAN)
Email: humanethics@rmit.edu.au
Tel: [61 3] 9925 4620

Notice of Approval

Date:	7 August 2020
Project number:	23125
Project title:	Identifying Aspects Contributing to Answer Utility in Question Answering
Risk classification:	Negligible/Low
Chief investigator:	Professor Mark Sanderson
Status:	Approved
Approval period:	From: 3/08/2020 To: 7/12/2020

The above application has been approved by the RMIT University CHEAN as it meets the requirements of the *National Statement on Ethical Conduct in Human Research* (NHMRC, 2007).

Terms of approval:

1. Responsibilities of chief investigator

It is the responsibility of the above chief investigator to ensure that all other investigators and staff on a project are aware of the terms of approval and to ensure that the project is conducted as approved by CHEAN. Approval is valid only whilst the chief investigator holds a position at RMIT University.

2. Amendments

Approval must be sought from CHEAN to amend any aspect of a project. To apply for an amendment, use the request for amendment form, which is available on the HREC website and submitted to the CHEAN secretary. Amendments must not be implemented without first gaining approval from CHEAN.

3. Adverse events

You should notify the CHEAN immediately (within 24 hours) of any serious or unanticipated adverse effects of their research on participants, and unforeseen events that might affect the ethical acceptability of the project.

4. Annual reports

Continued approval of this project is dependent on the submission of an annual report. Annual reports must be submitted by the anniversary of approval of the project for each full year of the project. If the project is of less than 12 months duration, then a final report only is required.

5. Final report

A final report must be provided within six months of the end of the project. CHEAN must be notified if the project is discontinued before the expected date of completion.

6. Monitoring

Projects may be subject to an audit or any other form of monitoring by the CHEAN at any time.

7. Retention and storage of data

The investigator is responsible for the storage and retention of original data according to the requirements of the *Australian Code for the Responsible Conduct of Research* (R22) and relevant RMIT policies.

8. Special conditions of approval

Nil.

In any future correspondence please quote the project number and project title above.

Yours faithfully,

Professor Falk Scholer

Chair, Science Engineering & Health

College Human Ethics Advisory Network

Cc Student investigator/s:

Other co-investigator/s:

Prof Falk Scholer, Dr Berkant Cambazoglu, Mrs Leila Tavakoli, Mrs Valeria Baranova,