Evaluating Robustness in Extractive Question Answering Systems with Unanswerable Queries

Noam Azulay **Daniel Samira** **Yuval Gorodissky** **Alon Neduva**
212988232 **316164417** **318963436** **316138221**

Ben-Gurion University of the Negev, Israel
Department of Software and Information Systems Engineering
{noamaz, samirada, yuvalgor, neduva}@post.bgu.ac.il

Abstract

This project is based on the article A Lightweight Method to Generate Unanswerable Questions in English [1]. We evaluate the methods described in this article, extending their application to various model architectures and both in-domain and out-of-domain tasks. Our experiments involve fine-tuning models using diverse training and testing datasets. We found that the Entity method, one of the proposed methods from the original paper, improves performance in in-domain tasks, while other methods have shown better performance in out-of-domain scenarios. Our study demonstrates the robustness of this lightweight approach and highlights the trade-offs between handling answerable and unanswerable questions across different architecture models. Our results and code are available at: https://github.com/yuvalgorodissky/computational_semantics_project.

1 Introduction

Natural Language Processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics. Natural Language Understanding (NLU), a subfield of NLP, involves the comprehension and interpretation of human language by machines, forming the backbone of advanced AI applications. In this project, we focus on Extractive Question Answering (EQA), a fundamental task in NLU.

1.1 Extractive Question Answering With Unanswerable Questions

Extractive Question Answering (EQA) systems are designed to pinpoint and retrieve the exact text segment from a provided input that answers a specific question. These systems, essential in the realm of NLP, excel in offering precise and succinct responses directly from the text.

A crucial enhancement in EQA is the inclusion of "no answer" scenarios—situations where the query does not have an explicit answer within the given text. This adaptation has introduced a more realistic and complex dimension to EQA tasks. In real-world applications, questions often arise without clear, extractable answers from a single text source. Thus, the ability of EQA systems to recognize and appropriately respond to such queries by indicating the absence of an answer is vital.

This capability not only tests the robustness and adaptability of EQA systems but also aligns more closely with human-like understanding and interaction. It addresses a challenge in developing more comprehensive QA systems that can effectively manage a broader spectrum of user inquiries, making them crucial for advanced NLP applications. The integration of "no answer" scenarios thus marks a significant step forward in making EQA systems more realistic tools for information retrieval.

1.2 Generating Unanswerable Questions

Most QA datasets primarily focus on answerable questions with answers found within the provided contexts. This approach leaves a gap in training QA systems to recognize when no answer is available based on the provided information. Generating unanswerable questions in English aims to challenge and improve the robustness of QA systems, which is crucial given their frequent occurrence in real-world scenarios.

1.3 Transformer Architectures

Recent advancements have propelled the development of AI models that are increasingly efficient and faster, thanks to innovations in transformer architecture. These developments have significantly enhanced the capabilities of QA systems, establishing new benchmarks for performance and effectiveness. Transformer models are typically classified into three primary architectures: encoder, decoder, and encoder-decoder as presented in Figure 1.

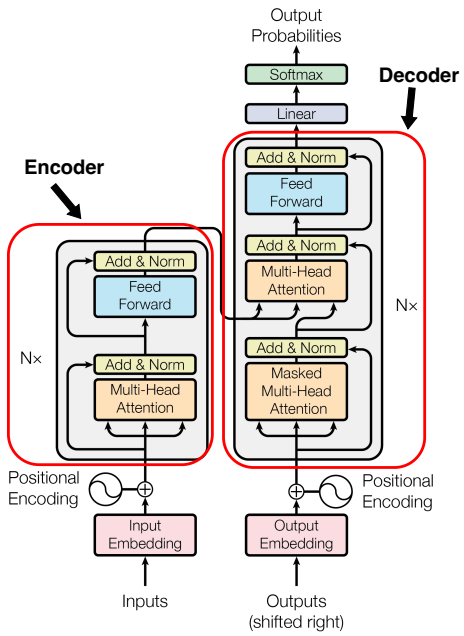


Figure 1: Schematic diagram of a transformer architecture highlighting the flow from input to output embeddings through the encoder and decoder structures [2].

1.3.1 Encoder

Encoder architectures are pivotal for tasks that require a deep understanding of input data. They excel in transforming textual inputs into dense vector representations that encapsulate the semantic essence of the text. This capability makes encoder models particularly effective for applications such as sentence classification and information retrieval, where comprehending the content is more critical than generating new text.

1.3.2 Decoder

Decoder models are designed to generate textual sequences from structured inputs. They play a crucial role in applications like text generation, where the ability to produce coherent, contextually relevant text from a given input or context is essential.

1.3.3 Encoder-Decoder

The encoder-decoder architecture synergies the abilities of both encoders and decoders. In this configuration, the encoder first interprets the input text and converts it into a meaningful representation. This representation is then utilized by the decoder to generate an appropriate output sequence.

1.4 Out-of-Domain Data

Out-of-domain data refers to data that comes from a different distribution than the data used during the model's training. Unlike in-domain data, which is similar to the training data, out-of-domain data poses a greater challenge as the model needs to generalize its knowledge to new and unseen scenarios. Evaluating models on out-of-domain tasks is essential for assessing their robustness and real-world applicability.

Research Questions:

- What is the performance of the methods across different model architectures, including encoder, decoder, and encoder-decoder models?
- How does the method perform on out-of-domain tasks?
- What is the trade-off between performance on answerable and unanswerable questions across different model architectures?

2 Related Work

In this section, we review the techniques employed to generate unanswerable questions, focusing on the transition from human-driven to automated processes and their implications for QA system development.

2.1 Generating Unanswerable Questions Methods

2.1.1 Human-Edited

Datasets such as SQuAD 2.0 [3], Natural Questions [4], and TydiQA [5] significantly contribute to QA research by including human-edited unanswerable questions designed to test systems' ability to determine when no answer is available based on the provided context. The generation of unanswerable questions by human editors, although crucial, is also highly resource-intensive in terms of both time and money. This underscores the need for developing automated methods that can efficiently replicate this process to reduce costs and accelerate dataset creation.

2.1.2 Automated Methods

In addition to human-edited techniques, automated methods such as UNANSQ [6] and CRQDA [7] have been developed to generate unanswerable questions. UNANSQ employs a model to transform answerable questions into unanswerable ones by adding plausible yet incorrect information. Similarly, CRQDA uses a sophisticated machine learning model to rephrase questions into forms that are contextually relevant but ultimately unanswerable.

While both methods mark significant advances in automating the generation of training data for QA systems, they require considerable computational resources and complex model training.

Another recent contribution is detailed in [1], which introduces a straightforward method for generating unanswerable questions in English. This paper proposes a lightweight method, consisting of fewer than 150 lines of Python code, emphasizing efficiency without the need for large, parameter-heavy models. The core strategies employed are:

- **Antonym Augmentation:** This method modifies answerable questions by replacing key words with their antonyms to ensure unanswerability.
- **Entity Augmentation:** This method alters questions by swapping named entities with others of the same type within the given context, thus preserving logical entity relationships and ensuring readability.

2.2 Dataset Selection

Our project utilizes three datasets in our experiments to provide a thorough and varied testing ground for the methods.

SQuAD 2.0 [3]: This dataset is a comprehensive extension of the Stanford Question Answering Dataset that features over 150,000 questions. It includes both answerable and unanswerable questions, making it particularly useful for testing the ability of models to not only retrieve answers but also recognize when no relevant information is available in the provided text.

TydiQA [5]: This dataset is a multilingual question-answering dataset covering 11 languages designed to assess the performance of QA systems across diverse linguistic features. It presents unique challenges due to structural, grammatical, and script variations in languages such as Arabic, Bengali, Finnish, Korean, and Russian. We utilize only the English portion of this dataset for our analysis.

ACE-whQA [8]: consists of three distinct subsets, each designed to test different aspects of question-answering systems:

- **ACE-whQA-has-ans**: This subset includes questions that have definitive answers, derived from the Automatic Content Extraction (ACE) program focused on event detection and entity recognition. It evaluates the model's ability to retrieve correct answers from complex informational contexts.
- **IDK Competitive**: This subset features unanswerable questions where the absence of an answer is not immediately apparent, challenging the model's capability to discern when information required to answer a question is missing in a competitive context.
- **IDK Non-Competitive**: Comprising unanswerable questions that are clearly lacking direct answers, this subset tests the model's efficiency in recognizing overtly absent information without misleading contexts.

3 Problem Definition

The goal of our project is to evaluate the robustness of the method described in the paper [1]. The original paper's method was validated only on encoder models and in-domain tasks. Our objective is to extend this evaluation to various models and architectures, including both encoder, decoder, and encoder-decoder models, and to assess the method's performance on out-of-domain tasks. We aim to broaden the current evaluation to encompass a variety of tasks and models, thereby deepening the understanding of the method's adaptability and efficiency across different scenarios.

4 Experimental Setup

In this section, we explain the taxonomy presented in Figure 2. The experiments include the following steps:

1. **Dataset Preparation:** Utilizing the dataset generated by the original method and the baseline datasets.
2. **Model Fine-Tuning:** Fine-tuning the FLAN-$T5_{Base}$ [9], Llama 3 [10], and $BERT_{large}$ [11] models (An encoder-decoder, decoder, encoder respectively). The fine-tuning process includes several sessions on different datasets. The models are fine-tuned using the SQuAD 2.0 [3] dataset. Other fine-tuning sessions are carried out on each baseline dataset—UNANSQ [6] and CRQDA [7], as well as on methods that are fine-tuned using the datasets from the original paper.
3. **In- and Out-of-Domain Evaluation:** Assess the models' performance on the in-domain

SQuAD 2.0 [3] and out-of-domain ACE-whQA [8] and TydiQA [5] datasets across models that were trained on different training sets.

4. **Robustness Analysis:** Evaluating and comparing the F1 score and EM (Exact Match) to determine how different architectures are influenced by various training data.

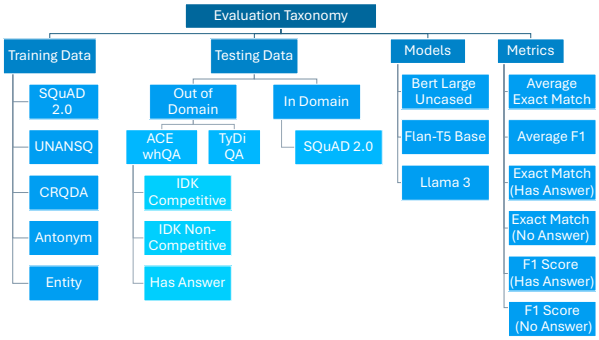


Figure 2: Our Evaluation Taxonomy outlining the methodologies and frameworks employed for generating and evaluating unanswerable questions.

5 Results and Discussion

The average F1 scores across all architecture models are demonstrated in Table 1 and in the heatmaps shown in Figure 4.

5.1 Performance Across Different Model Architectures

In the analysis of different model architectures, we observed distinct patterns of improvement and sensitivity to training methodologies.

Notably, the FLAN model demonstrated minimal variability in response to different training types compared to other models, suggesting it's performance is less influenced by the specific nature of training adjustments.

Across all examined methods, there was a consistent improvement over the baseline performance on the SQuAD 2 dataset. Furthermore, the training techniques of Entity and Antonym proposed in the [1] paper yielded superior results, particularly the Entity method, which closely aligned with the findings of the original study. These results were successfully replicated, confirming the robustness and applicability of the proposed methods across various model architectures. Both the BERT and FLAN models demonstrated improvements consistent with those reported in the article, reinforcing the stability of the method across in-domain tasks.

5.2 Domain Analysis

The results of our experiment support the conclusions drawn in the paper [1] regarding the performance of the BERT model in in-domain tasks. Specifically, the use of the Entity method improved the model's average F1 score compared to other tested methods. However, when extending these tests to out-of-domain tasks, our findings diverged. We discovered that multiple methods yielded the favorable outcomes across most out-of-domain tasks for BERT models but non of those were the Entity method. For example, in the TyDi QA dataset, UNANSQ led to improvements of 4.2 and 6.8 points over the Entity and Antonym methods, respectively, and an improvement of 6.2 points over the baseline. Furthermore, when compared to the baseline, UNANSQ showed substantial gains in the ACE-IDK Competitive and ACE-IDK Non-Competitive datasets, with at least 15.2 points in the Competitive and 19.1 points in the Non-Competitive setups. Similar results can be seen with the CRQDA training set as well. This suggests a reduction in the robustness of the methods proposed in the [1] paper when applied to out-of-domain contexts.

In contrast to the BERT model findings, the T5-Flan model, when trained using the Entity method, exhibited improvement in F1 scores for both in-domain and out-of-domain tasks. This method appears to be the most effective for this architecture based on our evaluations. Not only did it enhance performance consistently across different datasets, but it also showed a notable capacity to adapt to the varying demands of these tasks. This suggests that the Entity method may provide a robust training strategy for the T5-Flan model, optimizing it for a wide range of applications.

In the analysis of the Llama 3 model, the Antonym method demonstrated better and more balanced performance than the Entity method. Additionally, when the model was trained using the CRDQA method, it appeared slightly overfitted to unanswerable questions. This was evidenced by disproportionately high results on datasets that included only unanswerable questions compared to those containing only answerable questions.

Model	Train Set \ Test Set	SQuAD 2.0 - Dev	TyDi QA	AIC	AINC	AIHA
BERT$_{large}$	Baseline	79.36	29.29	2.80	6.09	54.08
	CRDQA	79.83	33.87	**64.00**	**93.08**	54.07
	UNANSQ	80.33	**35.48**	47.20	77.64	71.46
	Antonym	81.02	28.74	32	58.53	**78.49**
	Entity	**82.05**	31.29	18.8	28.04	75.93
FLAN$_{T5}$	Baseline	80.41	39.83	35.2	72.76	80.95
	CRDQA	80.56	41.21	36.80	70.32	79.34
	UNANSQ	80.36	40.36	34.40	70.32	80.45
	Antonym	80.85	40.16	30.40	65.85	**82.01**
	Entity	**80.99**	**41.65**	**38.4**	**74.79**	80.87
Llama$_3$	Baseline	79.50	37.50	50.8	62.19	52.32
	CRDQA	65.77	37.47	**83.2**	**88.61**	17.65
	UNANSQ	**80.68**	37.53	61.6	70.32	47.77
	Antonym	80.29	**37.63**	49.6	59.35	**65.09**
	Entity	78.92	37.36	54.8	67.88	50.93

Table 1: Average F1 Score comparison of models on all datasets. SQuAD 2.0 (in-domain) and CRDQA, UNANSQ, TyDi QA, ACE-IDK Competitive, ACE-IDK Non-Competitive, ACE-IDK Has Answer (out-of-domain). AIC = ACE-IDK Competitive, AINC = ACE-IDK Non-Competitive, AIHA = ACE-IDK Has Answer.

5.3 Trade-Off Between Answerable And Unanswerable

The results of F1 scores only on questions with answers and only on unanswerable questions for the BERT model are demonstrated in Figure 3. In the original article [1], the authors highlighted a trade-off when incorporating questions without answers into the training data on the SQuAD 2.0 development set, an in-domain task. While the model's ability to handle questions without answers improved, its performance on questions with answers deteriorated. This phenomenon was also evident in our results. For example, the BERT model's F1 score for answerable questions (at baseline conditions - BL) was 81.6, which was the second-highest score following the Antonym method, which achieved an F1 score of 81.9. In contrast, other methods showed a decrease in the F1 score for answerable questions, ranging from 3.5 to 7.4 points. Conversely, there was an increase in the F1 score for unanswerable questions, starting from an F1 score of 77.1 with the BL method. The scores were significantly lower compared to other models and training methods, with gaps of 8.9 and 9.3 points compared to the Entity and UNANSQ methods, respectively. These results were consistent in the BERT model tests and were similarly observed in the FLAN model, further supporting the hypothesis of a trade-off between handling answerable and unanswerable questions.

In the BERT model's performance on out-of-domain tasks, we observed two distinct outcomes. Similar to the in-domain results, in the TyDi QA dataset, there was an average decrease of 3 percent in the F1 score for answerable questions using the BL method compared to other methods. This decrease coincided with improvements in the F1 score for unanswerable questions, which varied between 1.5 and 19.2 points; the highest score achieved was 50.5 using the UNANSQ method. Surprisingly, we also noticed significant improvements in both the F1 score for answerable questions and unanswerable questions with other methods. in the ACE-whQA-has-ans dataset, peaking at 24.4 points improvement where the model employing the Antonym method achieved the highest F1 score of 78.5 for answerable questions. In the ACE-IDK Competitive and ACE-IDK Non-Competitive datasets, we also observed significant improvement in all methods, especially in the CRQDA method with an improvement 61.2 and 87.0 f1 score for unanswerable questions (competitive and non competitive respectively).

The trade-off observed between performance on answerable and unanswerable questions on out-of-domain tasks seems less pronounced in the T5-Flan model. This observation is consistent with our broader analysis, which indicated that the FLAN

model generally shows minimal variability when responding to different training methodologies even with the baseline.

In general, the F1 scores from the Llama 3 model exhibit instability, highlighting a trade-off between answerable and unanswerable questions. While some methods enhance performance on in-domain tasks, they falter on out-of-domain unanswerable questions. For instance, the model employing the CRQDA method excels in out-of-domain scenarios but struggles with answerable questions. Conversely, other methods demonstrate a more balanced performance, effectively bridging the gap between answerable and unanswerable questions.

6 Discussions

In this project, we aimed to extend the evaluation of the method for generating unanswerable questions proposed in the article [1] to a variety of model architectures and both in-domain and out-of-domain tasks.

Our study evaluated the effectiveness of various training methods across different model architectures, including encoder (BERT), decoder (Llama 3), and encoder-decoder (FLAN) models. The results demonstrated that the FLAN model exhibited minimal variability in response to different training methodologies, suggesting a high degree of stability and robustness in its performance. Both the BERT and FLAN models showed enhancements in domain-specific tasks, with consistent improvements over baseline performance on the SQuAD 2 dataset. The Entity and Antonym training methods, as proposed in prior studies, generally yielded the highest results, especially the Entity method, which aligned closely with original findings. This consistency underscores the reliability of these methods across various architectures for in-domain tasks.

Out-of-Domain Evaluation: Our results indicated a divergence in the effectiveness of the methods. The BERT model, while showing improvements with the Entity method for in-domain tasks, did not sustain these gains in out-of-domain contexts. Instead, training datasets like UNANSQ and CRDQA showed more substantial improvements across datasets such as TyDi QA and ACE-IDK, highlighting a reduction in the robustness of the Entity method in out-of-domain scenarios. In contrast, the FLAN model maintained robust performance with the Entity method across both in-domain and out-of-domain tasks, suggesting its adaptability to varied datasets. The Llama 3 model, however, showed better performance with the Antonym method and exhibited overfitting issues with the CRDQA method, particularly on datasets with only unanswerable questions.

Trade-Off Analysis: Our analysis identified a trade-off between the performance on answerable and unanswerable questions across different model architectures. For the BERT model, we observed the trade-off in the in-domain task as described in the original paper and in the TyDi QA test set, but no trade-off was observed in the ACE-IDK test set. In fact, improvements in both types of questions were demonstrated there. The FLAN model, however, showed minimal trade-off, indicating balanced performance with stable F1 scores across both types of questions. The Llama 3 model highlighted instability, with some methods like CRDQA excelling in out-of-domain unanswerable questions but struggling with answerable questions.

6.1 Contributions

Our study confirmed the robustness and applicability of the lightweight method for generating unanswerable questions across different model architectures and datasets. By broadening the evaluation to include various models and tasks, we provided deeper insights into the method's adaptability and efficiency. Our findings support the method's potential to enhance the realism and comprehensiveness of EQA systems, making them better tools for real-world information retrieval.

6.2 Future Work

Future research could explore further optimization of training techniques to minimize the trade-off between answerable and unanswerable question performance. Additionally, extending the evaluation to include diverse datasets and real-world applications would provide a more comprehensive understanding of the method's capabilities and limitations.

References

[1] Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow. A lightweight method to generate unanswerable questions in english. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7349–7360, 2023. URL https://aclanthology.org/2023.findings-emnlp.491.pdf.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[3] 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://aclanthology.org/P18-2124.

[4] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466, 2019.

[5] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317. URL https://aclanthology.org/2020.tacl-1.30.

[6] Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4238–4248, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1415. URL https://aclanthology.org/P19-1415.

[7] Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Jiancheng Lv, Nan Duan, and Ming Zhou. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5798–5810, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.467. URL https://aclanthology.org/2020.emnlp-main.467.

[8] Elior Sulem, Jamaal Hay, and Dan Roth. Do We Know What We Don't Know? Studying Unanswerable Questions beyond SQuAD 2.0. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://cogcomp.seas.upenn.edu/papers/SulemHaRo21.pdf.

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Appendices

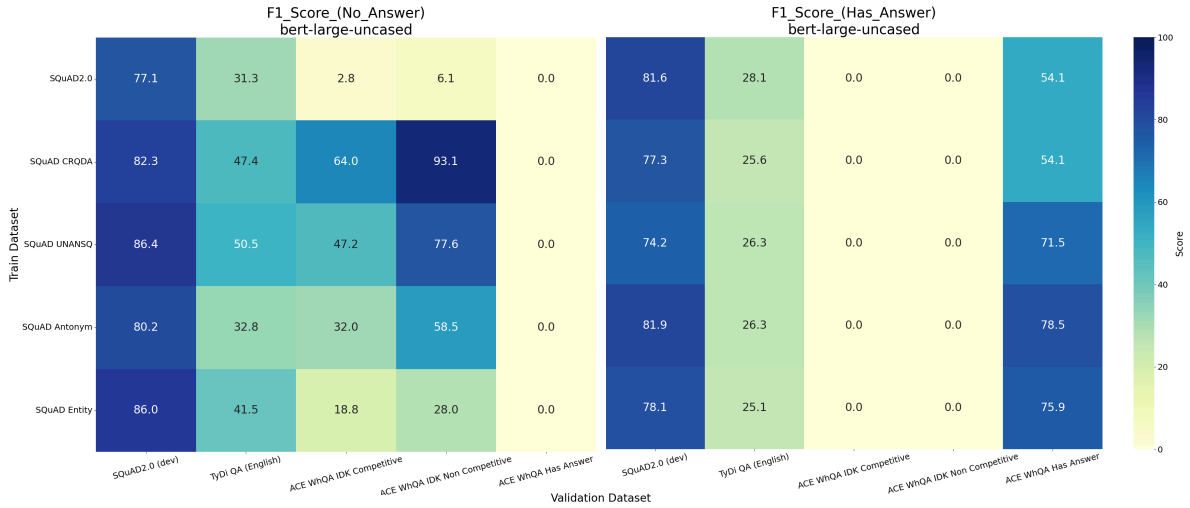


Figure 3: Comparison of average F1 scores for BERT with Answer and No Answer.

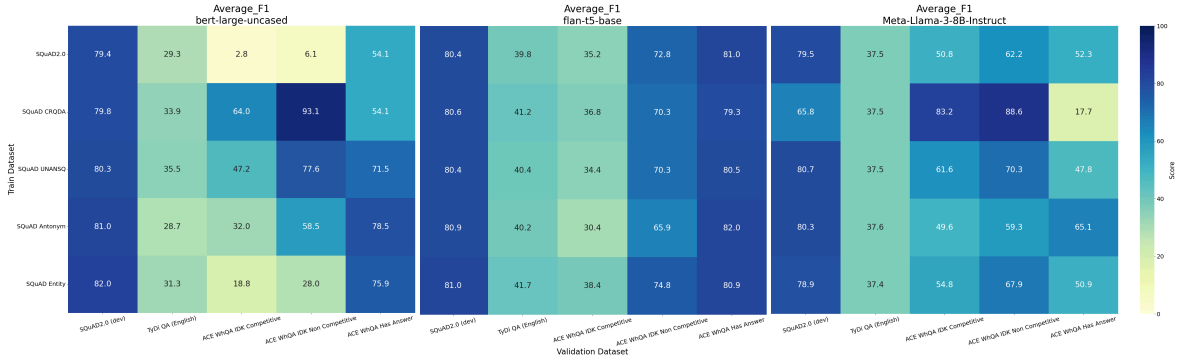
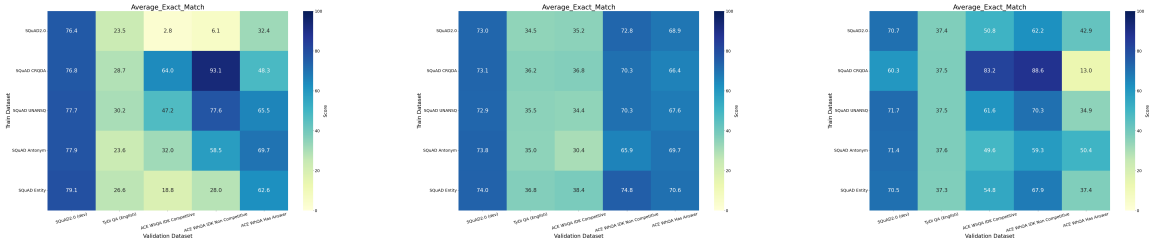
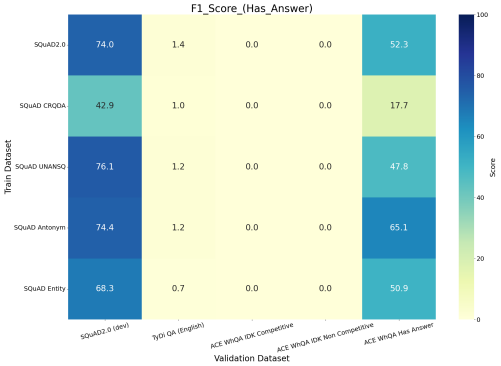


Figure 4: Comparison of average F1 scores for different models.

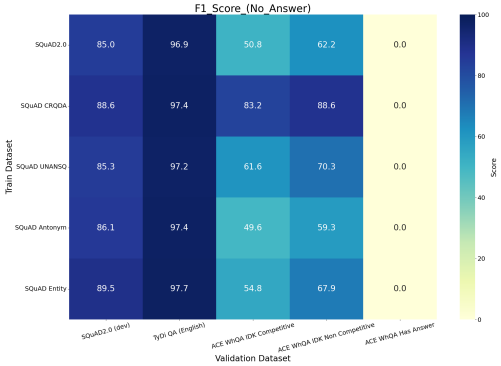


(a) BERT Large Uncased. (b) FLAN-T5 Base. (c) Meta LLaMA 3 8B Instruct.

Figure 5: Comparison of average Exact Match scores for different models.

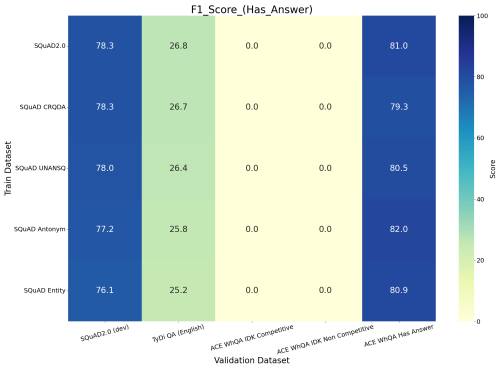


(a) Meta-Llama-3-8B-Instruct with Answer.

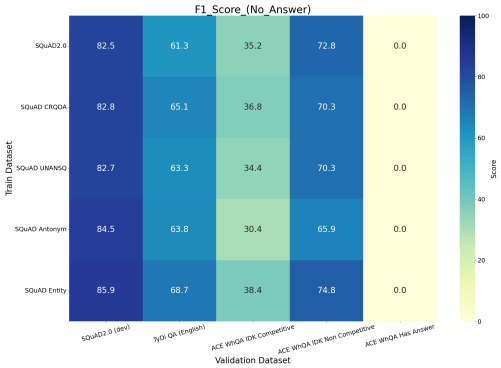


(b) Meta-Llama-3-8B-Instruct with No Answer.

Figure 6: Comparison of average F1 scores for Meta-Llama-3-8B-Instruct with Answer and No Answer.



(a) FLAN-T5 Base with Answer.



(b) FLAN-T5 Base with No Answer.

Figure 7: Comparison of average F1 scores for FLAN-T5 Base with Answer and No Answer.