How Do We Answer Complex Questions: Discourse Structure of Long-form Answers

Fangyuan Xu[♢] **Junyi Jessy Li**[♡] **Eunsol Choi**[♢]
[♢]Department of Computer Science
[♡]Department of Linguistics
The University of Texas at Austin
{fangyuan, jessy, eunsol}@utexas.edu

Abstract

Long-form answers, consisting of multiple sentences, can provide nuanced and comprehensive answers to a broader set of questions. To better understand this complex and understudied task, we study the functional structure of long-form answers collected from three datasets, ELI5 (Fan et al., 2019), WebGPT (Nakano et al., 2021) and Natural Questions (Kwiatkowski et al., 2019). Our main goal is to understand how humans organize information to craft complex answers. We develop an ontology of six sentence-level functional roles for long-form answers, and annotate 3.9k sentences in 640 answer paragraphs. Different answer collection methods manifest in different discourse structures. We further analyze model-generated answers – finding that annotators agree less with each other when annotating model-generated answers compared to annotating human-written answers. Our annotated data enables training a strong classifier that can be used for automatic analysis. We hope our work can inspire future research on discourse-level modeling and evaluation of long-form QA systems.[1]

1 Introduction

While many information seeking questions can be answered by a short text span, requiring a short span answer significantly limits the types of questions that can be addressed as well as the extent of information that can be conveyed. Recent work (Fan et al., 2019; Krishna et al., 2021; Nakano et al., 2021) explored long-form answers, where answers are free-form texts consisting of multiple sentences. Such long-form answers provide flexible space where the answerer can provide a nuanced answer, incorporating their confidence and sources of their knowledge. Thus the answer sentences form a *discourse* where the answerers provide information, hedge, explain, provide examples, point to other sources, and more; these elements need to be structured and organized coherently.

We take a linguistically informed approach to understand the structure of long-form answers, designing six communicative *functions* of sentences in long-form answers (which we call **roles**).[2] Our framework combines functional structures with the notion of information salience by designating a role for sentences that convey the main message of an answer. Other roles include signaling the organization of the answer, directly answering the question, giving an example, providing background information, and so on. About a half of the sentences in long-form answers we study serve roles other than providing an answer to the question.

We collect discourse annotations on three long-form question answering (LFQA) datasets, ELI5 (Fan et al., 2019), WebGPT (Nakano et al., 2021) and Natural Questions (NQ) (Kwiatkowski et al., 2019). Figure 1 contains an example annotation on each dataset. While all three contain paragraph-length answers needed for complex queries, they are collected in distinct manners – answers in ELI5 are written by Reddit users; answers in WebGPT are written by annotators who searched documents on a web interface and heavily quoted those documents to form an answer, and answers in NQ are pre-existing paragraphs from Wikipedia corpus. We collect three-way annotations for 3.9k sentences (∼700 question-answer pairs across three datasets). We also annotate a small number of model-generated answers from a recent long-form question answering (LFQA) system (Krishna et al., 2021) and provide rich analysis of their discourse structure.

In all three datasets, we observe appearance of most proposed functional roles, but with different proportions. Answers in ELI5 contains more examples and elaborations, while answers extracted

[1]Our data, code and datasheet are available at https://github.com/utcsnlp/lfqa_discourse.

[2]Functional structures have been studied in various other domains (discussed in Sections 2 and 7).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3556 - 3572

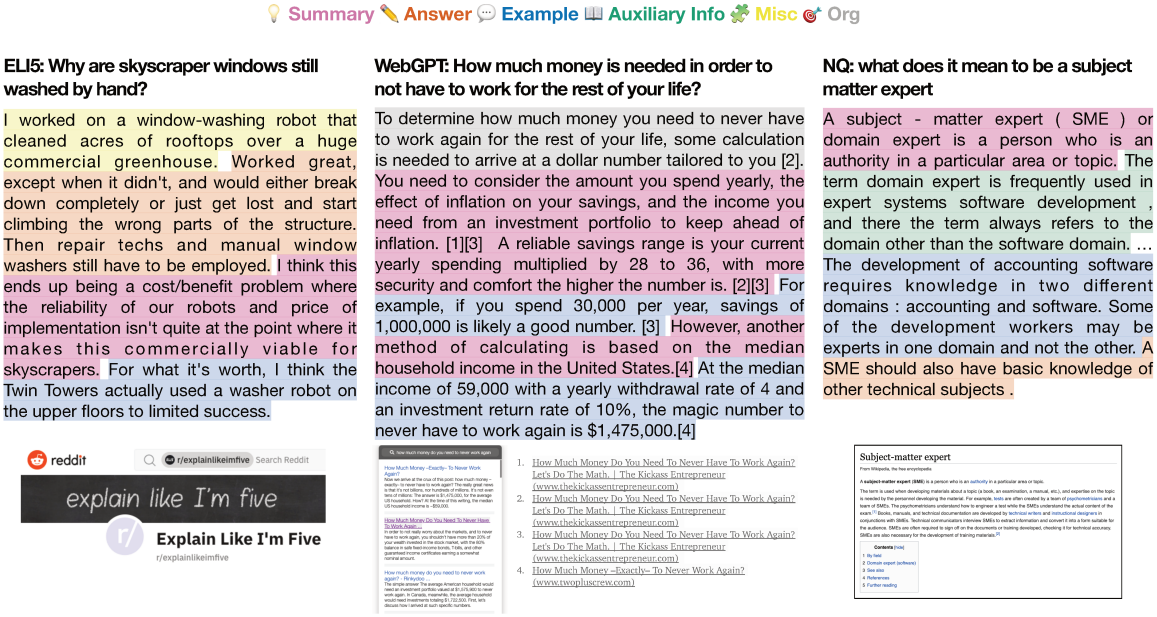

Figure 1: Long-form answers from ELI5, WebGPT and NQ dataset. Each sentence in the answer is annotated with a sentence-level functional role from our ontology, described in Section 2.

from Wikipedia passages (NQ) contain more auxiliary information. Analyzing a subset of ELI5 and WebGPT, we also identify a big gap in lexical overlap between long-form answer and evidence passages across all functional roles. Lastly, we found that human agreement of the discourse roles of model-generated answers are much lower than human-written ones, reflecting the difficulty for humans to process model-generated answers.

With the data collected, we present a competitive role classifier, which performs on par with human when trained with our annotated data and can be used for automatic discourse analysis. We further envision using functional roles for controllable long-form generations, concise answer generation, and improved evaluation metrics for LFQA.

2 Defining Answer Discourse Structure

We study the discourse structure of long-form answers based on *functional roles* of sentences in the paragraph. Functional structures characterize the communicative role a linguistic unit plays; as such, they vary across genres as the goals of communication also vary. In scientific or technical articles, these roles can be *background, method, findings* (Kircz, 1991; Liddy, 1991; Mizuta et al., 2006), while in news, they can be *main event* or *anecdotes* (Van Dijk, 2013; Choubey et al., 2020).

These structures are related to, though distinct from, coherence discourse structures (Hobbs, 1985). The latter characterizes how each unit (e.g., adjacent clauses or sentences) *relates* to others through semantic relations such as temporal, causal, etc.; such structures can be trees that hierarchically relate adjacent units (Mann and Thompson, 1988) or graphs (Lascarides and Asher, 2008). In contrast, functional roles describe how information is organized to serve the communication goal, in our case, providing the answer.

We developed our ontology by examining long-form answers in online community forums (subreddit *Explain Like I'm Five* (ELI5)) and Wikipedia passages, hence answers derived from different domains (e.g., textbooks) can contain roles beyond our ontology. We describe our six sentence-level discourse roles for long-form answers here:

Answer-Summary (Sum), Answer (Ans). An answer sentence directly addresses the question. Here we distinguish between the the main content of the answer (henceforth *answer summary*) vs. sentences which explain or elaborate on the summary. The summaries play a more salient role than non-summary answer sentences, and can often suffice by themselves as the answer to the question. This is akin to argumentation structure that hierarchically arranges main claims and supporting

arguments (Peldszus and Stede, 2013), and news structure that differentiates between main vs. supporting events (Van Dijk, 2013).

Organizational sentences (Org.) Rather than conveying information of the answer, the major role of an organizational sentence is to inform the reader how the answer will be structured. We found two main types of such sentences; the first signals an upcoming set of items of parallel importance:

> **[A]**: **There are a few reasons candidates with "no chance" to win keep running**. 1) They enjoy campaigning[...]

The other type indicates that part of the answer is upcoming amidst an established flow; in the example below, the answerer used a hypophora:

> **[A]**: It might actually be a mosquito bite. I find the odd mosquito in my house in the winter from time to time, and I'm in Canada.[...] **So why does it happen more often when you shower?** It's largely because [...]

Examples (Ex.) Often people provide examples in answers; these are linguistically distinct from other answer sentences in the sense that they are more specific towards a particular entity, concept, or situation. This pattern of language specificity can also be found in example-related discourse relations (Louis and Nenkova, 2011; Li and Nenkova, 2015), or through entity instantiation (MacKinlay and Markert, 2011):

> **[Q]**: What is it about electricity that kills you?
> **[A]**: [...] **For example, static electricity consists of tens of thousands of volts, but basically no amps.** [...]

We found that examples in human answers are often not signaled explicitly, and often contain hypothetical situations:

> **[Q]**: Were major news outlets established with political bias or was it formed over time?
> **[A]**: [...]This is impossible due to the problem of "anchoring." **Consider a world where people on the right want the tax rate to be 1% lower and people on the left want the tax rate to be 1% higher**[...]

Auxiliary information (Aux.) These sentences provide information that are related to what is discussed in the answer, but not asked in the question. It could be background knowledge that the answerer deemed necessary or helpful, e.g.,

> **[Q]**: Why is it better to use cloning software instead of just copying and pasting the entire drive?
> **[A]**: **When you install an operating system, it sets up what's called a master file table, which [...] are important for the OS to work properly.** [...] Simply copy-pasting files doesn't copy either of these, meaning if you want to back up an OS installation you should clone the disk instead.

or related content that extends the question, e.g.,

> **[Q]**: what is the difference between mandi and kabsa?
> **[A]**: [...] A popular way of preparing meat is called mandi. [...] **Another way of preparing and serving meat for kabsa is mathbi , where seasoned meat is grilled on flat stones that are placed on top of burning embers.**

Notably, the removal of auxiliary information would still leave the answer itself intact.

Miscellaneous (Misc.) We observe various roles that, although less frequent, show up consistently in human answers. We group them into a *miscellaneous* role and list them below.

(a) Some sentences specify the limitation of the answer by narrowing down the scope of the answer to an open-ended question.

> **[Q]**: Why are there such drastic differences in salaries between different countries?
> **[A]**: I'm going to avoid discussing service industries, because[...] **I'm mostly talking tech.** [...]

(b) Some sentences state where the answer came from and thus put the answer into context.

> **[Q]**: Why Does a thermostat require the user to switch between heat and cool modes, as opposed to just setting the desired temperature?
> **[A]**: **The person who installed my heat pump (which has all three modes) explained this to me.** [...]

(c) Some sentences point to other resources that might contain the answers.

> **[Q]**: Why did Catholicism embrace celibacy and why did Protestantism reject it?
> **[A]**: **/r/askhistorians has a few excellent discussions about this.** [...]

(d) Answerers also express sentiment towards other responses or the question itself.

> **[Q]**: Why did Catholicism embrace celibacy and why did Protestantism reject it?
> **[A]**: **Good God, the amount of misinformation upvoted is hurting.** [...]

> **[Q]**: Could you Explain Schrödinger's Cat to me LI5?
> **[A]**: [...] **It's a pretty cool thought experiment, but it doesn't mean too much in our everyday lives.**

As our ontology does not provide an exhaustive list of the functional roles, we instructed our annotators to annotate other roles not covered by our ontology as Miscellaneous as well.

3 Data and Annotation

3.1 Source Datasets

We randomly sample examples from three LFQA datasets and filter answers with more than 15 sentences and those with less than 3 sentences.[3] We briefly describe each dataset below.[4]

ELI5 / ELI5-model ELI5 consists of QA pairs where the questions and answers are retrieved from the subreddit `r/explainlikeimfive`. The answers in ELI5 are of varying quality and style. While the original dataset consists of (question, answer) pairs, recent benchmark (Petroni et al., 2021) annotated a subset of examples with relevant Wikipedia paragraphs, which we used for analysis in Section 4. In addition to answers in the original datasets, we annotate a small number of model-generated answers from Krishna et al. (2021) (we refer this set as ELI5-model), a state-of-the art LFQA system on ELI5.

WebGPT Nakano et al. (2021) presented a new LFQA dataset and model; with the goal of building a model that can search and navigate the web to compose a long-form answer. While they reuse questions from ELI5, they newly collect answers from trained human annotators who were instructed to first search for related documents using a search engine and then construct the answers with reference to those documents. The collected data (denoted as "human demonstration" consisting of question, answer, a set of evidence documents, and mapping from the answer to the evidence document) are used to finetune GPT-3 (Brown et al., 2020) to generate long-form answers.

Natural Questions (NQ) NQ contains questions from Google search queries, which is paired with a relevant Wikipedia article and an answer in the article if the article answers the question. They annotate paragraph-level answer as well as short span answer inside the paragraph answer if it exists. In open retrieval QA, researchers (Lee et al., 2019) filtered questions with paragraph level answers for its difficulty of evaluation and only look at questions with short span answer.

We create a filtered set of NQ that focuses on paragraph-level answers containing complex queries.[5] While many NQ questions can be answered with a short entity (e.g., *how many episodes in season 2 breaking bad?*), many others questions require paragraph length answer (e.g., *what does the word china mean in chinese?*). This provides a complementary view compared to the other two datasets, as the answers are not *written specifically* for the questions but harvested from pre-written Wikipedia paragraphs. Thus, this simulates scenarios where model retrieves paragraphs instead of generating them.

3.2 Annotation Process

We have a two-stage annotation process: annotators first determine the validity of the QA pair, and proceed to discourse annotation only if they consider the QA pair valid. We define the QA pair as valid if (1) the question is interpretable, (2) the question does not have presuppositions rejected by the answer, (3) the question does not contain more than one sub-question, and (4) the proposed answer properly addresses the question. Examples of the invalid QA pair identified are in A.1.[6]

We collect the first stage annotation from US-based crowdsource workers on Amazon Mechanical Turk and second stage annotation from undergraduate students majoring in linguistics, who are native speakers in English.[7] A total of 29 crowdworker participated in our task, and six undergraduates annotated roles for a subset of QA pairs annotated as valid by crowdworkers. We first qualified and then provided training materials to both groups of annotators. The annotation guideline and interface can be found in A.4. We paid crowd workers $0.5 per example, and our undergraduate annotators $13 / hour. More details of data collection can be found in our datasheet.

3.3 Data Statistics

Table 1 presents the statistics of our annotated data. We collected validity annotations for 1.5K exam-

[3] We used Stanza (Qi et al., 2020) to split long-form answers into sentences. This process removes 42%, 28% and 34% from ELI5, WebGPT and NQ respectively.

[4] Our data is sourced from the validation split of ELI5 from the KILT (Petroni et al., 2021) benchmark, the testing portion from WebGPT (their samples are publicly hosted at `https://openaipublic.blob.core.windows.net/webgpt-answer-viewer/index.html`, which answers questions from the ELI5 test set), and the validation split from Natural Questions.

[5] Implementation details are in A.3. We also release these questions in our github repository.

[6] The categories are not mutually exclusive, and we let annotators to pick any of them when an example belongs to multiple categories.

[7] Initially, we aimed to collect all data from crowdsourcing, but during our pilot we found that it is challenging for crowd worker to make role assignment.

Data	Validity	Role	Length
ELI5	1,035 (6,575)	411 (2,670)	6 (126)
ELI5-model	193 (1,839)	115 (1,080)	10 (210)
WebGPT	100 (562)	98 (551)	6 (131)
NQ	263 (1,404)	131 (695)	5 (139)
Total	1,591 (10,380)	755 (4,996)	7 (139)

Table 1: Data Statistics. For validity and role, the first number in each cell corresponds to the number of long-form answers, and the second number represents the number of sentences. For length, the first number corresponds to the average number of sentences and the second represents the number of words.

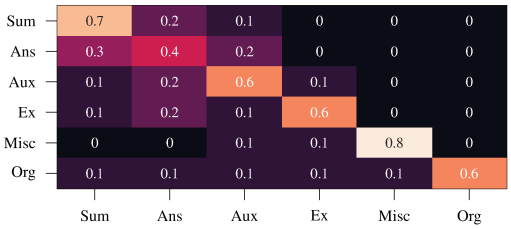

Figure 2: Confusion matrix of role annotations.

ples and role annotations for about half of them. As our tasks are complex and somewhat subjective, we collected three way annotations. We consider a QA pair valid if all annotated it as valid, and invalid if more than two annotated it as invalid. If two annotators considered valid, we collect one additional annotation and consider it valid if and only if the additional annotator marked it as valid.[8] We consider the majority role (i.e. chosen by two or more than two annotators) as the gold label. When all annotators chose different roles, they resolved the disagreement through adjudication. We report inter-annotator agreement before the adjudication.

Inter-annotator Agreement We find modest to high agreement for both annotation tasks: For crowdworkers, Fleiss Kappa was 0.51 for validity annotation. For student annotators, Fleiss Kappa was 0.44 for role annotation. Figure 2 shows the confusion matrix between pairs of annotations, with the numbers normalized by row and averaged across pairs of annotators. We observe frequent confusion between roles denoting different levels of information salience —Answer vs. Answer-Summary, and Answer vs. Auxiliary Information, reflecting the nuance and subjectivity in judging what information is necessary to answer a complicated question. Examples can be found in A.2.

[8]The Fleiss kappa for agreement improves to 0.70 after this re-annotation process.

Reason	NQ	ELI5	WebGPT
No valid answer	15%	10%	1%
Nonsensical question	3%	1%	0%
Multiple questions	9%	4%	1%
Rejected presupposition	2%	10%	0%
Total	23%	19%	2%

Table 2: Different reasons for invalid question answer pairs and their frequency in the three datasets.

4 Discourse Analysis of Long-form Answers

With our annotated data, we study the differences between the three types of long-form answers, namely answers provided by users in online community (ELI5), answers written by trained annotators through web search (WebGPT), and answers identified in Wikipedia passages (NQ).

Q/A Validity Table 2 summarizes the portion of valid answers in the three datasets and the distribution of invalid reasons. NQ has the highest rate of invalid answer (15%). Upon manual inspection, we find that passages from Wikipedia written independently of the question often only partially address complex questions. This demonstrates the limitation of a fully extractive approach. Around 10% of the answers from ELI5 reject presupposition in the question, which is a common phenomena in information-seeking questions (Kim et al., 2021). WebGPT boasts the lowest invalid rate, showing the high quality of their collected answers.

Role Distribution We study the distribution of roles in three datasets (Table 3). NQ shows the highest proportion of auxiliary information, as the paragraphs are written independent of the questions. In contrast, ELI5 contains more answer sentences and examples which provide explanation. Both ELI5 and WebGPT contain organizational sentences, demonstrating that it is commonly used when answerers assemble answers that cover more than one aspects. In all datasets, around half of the sentences serve roles other than directly answering the questions, such as providing auxiliary information or giving an example, which reflects the wide spectrum of information presented in a long-form answer. Relatively few sentences (less than 10%) are marked as miscellaneous, showing a high coverage of our ontology in the three LFQA datasets we investigated. Compared to ELI5, both WebGPT and NQ answers contain very little miscellaneous

Data	# of Annotated Sentences	Role					
		Answer	Summary	Auxiliary	Example	Org	Misc
ELI5	2670	30%	28%	18%	13%	1%	10%
WebGPT	551	28%	35%	26%	8%	3%	0%
NQ	695	21%	35%	39%	5%	0.4%	0.1%
Total	3916	28%	30%	11%	23%	1%	7%

Table 3: Sentence-level role distribution. The first column represent the total number of the annotated answer sentences. The remaining column represents the proportion of each role in respective datasets.

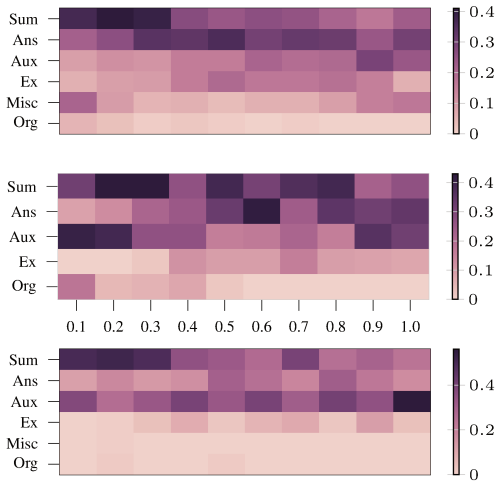

Figure 3: Heatmap of role distribution by the relative position in the answer paragraph in the three datasets, from top to bottom: ELI5, WebGPT, NQ.

sentences. This is partially because both datasets are more extractive and less personal, without sentences which serve the role of various kinds of communication from answerers to question askers (e.g. expressing sentiments, pointing to other resources) that are commonly seen in online community forum.

Discourse Structure Figure 3 presents the distribution of each role per its relative location in the answer. Despite the significant differences in the proportion of different discourse roles, the positioning of the roles is similar across the datasets. Answer summary and organizational sentences typically locate at the beginning of the paragraph, examples and answers often in the middle, with an increasing portion of auxiliary information towards the end. The sentences belonging to miscellaneous role frequently position at the beginning or the end of the paragraph, instead of intervening in the middle. WebGPT contains a higher portion of auxiliary information locating at the beginning of the passage, followed by the answer summary sentences.

Answer Extractiveness One important aspect for long-form answer is whether the answer can be attributed to an external evidence document. While answers from NQ are directly extracted from Wikipedia passages, both ELI5 and WebGPT are written specifically for the question. To help with verification, both datasets provide evidence documents paired with the answer, and yet there are design differences between the two. Answerer (annotators) of WebGPT were instructed to answer the question *based on* the evidence documents returned by a search engine, while answers from ELI5 were written first *independently* and later paired with relevant Wikipedia passages (Petroni et al., 2021).

We found that such difference leads to different level of extractiveness of the answer, by calculating sentence-level lexical overlap (after removing stopwords) with the evidence document. Overall, WebGPT answers exhibit **more** lexical overlap (unigram: 0.64, bigram: 0.36) with evidence document than ELI5 answers (unigram: 0.09, bigram: 0.01). Answer sentences with different roles also exhibit different levels of extractiveness (detailed role-level overlap can be found in Table 8 in the appendix). For ELI5 answers, sentences belonging to answer and summary roles have the highest overlap while example, auxiliary information and miscellaneous sentences are less *grounded* to external sources. For WebGPT, organizational sentences are the least extractive among all the roles.

5 Discourse Structure of Model-generated Answers

Having analyzed discourse roles of human-written long-form answers, we investigate the discourse structure of model-generated answers. This will allow us to quantitatively study the difference in terms of discourse structure across gold and generated answers, which we hope will cast insights to the linguistic quality of system outputs.

Systems We study model-generated answers from a state-of-the-art LFQA system (Krishna et al., 2021).[9] Their model uses passage retriever (Guu et al., 2020), and generates answers based on the retrieved passage with a routing transformer model (Roy et al., 2021).

Q/A Validity We collect validity annotation on 193 model-generated answers, and 78 are considered invalid, significantly higher ratio than that of human written answers (Table 2). The Fleiss's kappa of QA pair validity is 0.26 (and 0.61 after collecting one more label), substantially lower than the agreement on human written answers (0.51, 0.70) while annotated by the same set of annotators. Detailed distribution of invalid reason annotated can be found in Table 9. Despite the low agreement, 60 of them are marked as "No valid answer" by at least two annotators. The flaw of automatic measures was also pointed out by prior work (Krishna et al., 2021), which compares ROUGE between human-written and model-generated answers. Our study reiterates that the generated answers received high ROUGE score *without* answering the question.

Roles We proceed to collect sentence-level role annotations on 115 valid generated long-form answers following the same annotation setup in Section 3, and hence our annotators were not asked to evaluate the *correctness* or the *quality* of the answers (e.g. whether the generated example makes sense), focusing on the functional roles of sentences only. We found that the annotators *disagree* substantially more as compared to the human-written answers, with a Fleiss kappa of 0.31 (vs. 0.45 for human-written answers), suggesting that the discourse structure of model-generated answers are *less clear*, even to our trained annotators.

The answer role distribution of model-generated answers is very different from that of the human written answers (Figure 4). The generated answers contain more sentences which provide auxiliary information, and fewer summary sentences. This suggests that model-generated answers contain a higher portion of information tangentially related to what is asked in the question. Model-generated answers also contain fewer example and miscellaneous sentences. Examples of annotated model-generated answer can be found in Table 10.

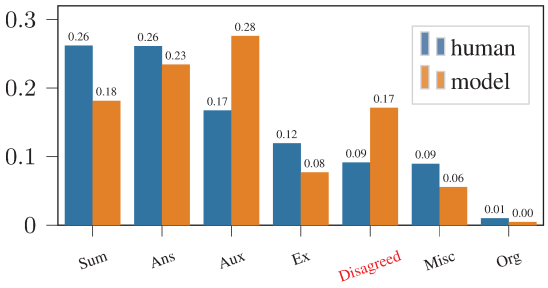

Figure 4: Annotated role distribution of model generated v.s. human written answers for ELI5 dataset, denoted by % sentence. We plot the majority role before adjudication and include those without a a majority role as "Disagreed".

Overall, our results suggest that machine-generated long form answers are different from human-written answers, and judging their discourse structure is nontrivial for human annotators, resulting in lower agreement. Recent study (Karpinska et al., 2021) also showed that expert annotators showed lower agreement and took longer time to evaluate the coherence of story generated from large-scale language model.

6 Automatic Discourse Analysis

We study how models can identify the discourse role for each sentence in long-form answer in a valid QA pair.[10] Such a model can be beneficial for large-scale automatic analysis.

6.1 Experimental Settings

Task and Data Given a question q and its long-form answer consisting of sentences $s_1, s_2 ... s_n$, the goal is to assign each answer sentence s_i one of the six roles defined in Section 2. As we have annotated more examples from ELI5 dataset (411 answer paragraphs compared to around 100 paragraphs in other three datasets (WebGPT, NQ and ELI5-model)), we randomly split the ELI5 long-form answers into train, validation and test sets with a 70%/15%/15% ratio, and train the model on the training portion. We use all other annotated datasets for evaluation only. For model-generated answers, we filtered 185 out of 1080 sentences where model-generated sentences do not have a majority role. This setup also allows us to study domain transfer of role classification model.

[9] We sampled from four different model configurations reported in their paper, i.e. combination of nucleus sampling threshold p={0.6, 0.9}, and generation conditioning on {predicted, random} passages. The answers we annotated achieved a ROUGE-L of 23.19, higher than that of human-written answers on the same set of questions (21.28).

[10] We do not automatically classify QA pair validity, which requires in-depth world knowledge and goes beyond the scope of our study.

System	Acc	Match	Macro-F1	Ans	Sum	Aux	Ex	Org	Msc
Majority	0.29	0.44	0.07	0	0.45	0	0	0	0
Summary-lead	0.36	0.56	0.15	0.44	0.46	0	0	0	0
RoBERTa	0.46	0.65	0.43	0.41	0.54	0.31	0.43	0.22	0.61
T5-base	0.48	0.67	0.45	0.44	0.46	0.35	0.52	0.06	0.86
T5-large	**0.54**	**0.75**	**0.55**	**0.49**	**0.55**	**0.46**	**0.59**	**0.44**	**0.79**
Human (l)	0.55	0.73	0.52	0.45	0.66	0.44	0.57	0.29	0.74
Human (u)	0.76	1.00	0.74	0.71	0.82	0.69	0.77	0.56	0.86

Table 4: Role identification results on test split of ELI5 dataset.

System	Acc			Match			Macro-F1		
	WebGPT	NQ	ELI5-Model	WebGPT	NQ	ELI5-Model	WebGPT	NQ	ELI5-Model
Majority	0.35	0.35	0.22	0.56	0.50	0.35	0.09	0.09	0.06
Summary-lead	0.35	0.34	0.35	0.54	0.56	0.53	0.15	0.15	0.16
RoBERTa	0.39	0.45	0.45	0.60	0.64	0.62	0.32	0.33	0.39
T5-base	0.43	0.44	0.43	0.65	0.61	0.58	0.38	0.35	0.42
T5-large	**0.48**	**0.45**	**0.46**	**0.70**	**0.64**	**0.64**	**0.46**	**0.40**	**0.48**
Human (l)	0.53	0.59	0.57	0.71	0.75	0.78	0.45	0.43	0.58
Human (u)	0.73	0.78	0.78	1.00	1.00	1.00	0.61	0.66	0.78

Table 5: Role identification results on out-of-domain datasets. Per-role performances are in Table 12 in the appendix.

Metrics We report accuracy with respect to the majority role label (or adjudicated one, if majority doesn't exist) (**Acc**), match on any label from three annotators (**Match**), **F1** score for each role and their macro average score **Macro-F1**.

6.2 Models

Lower bounds We present two simple baselines to provide lower bounds: (1) Majority: We predict the most frequent labels in the training data: *Answer-Summary*. (2) Summary-lead: We predict first two sentences as *Answer-Summary*, and the rest of the sentences as *Answer*.

Classification Models This baseline classifies each sentence independently. We use the [CLS] token from RoBERTa-Large model (Liu et al., 2019) which encodes [$question$ <q> ans_1 ... <start> ans_i <end> ...], where ans_i is the i^{th} sentence in the answer. The training batch size is set to 64, with the initial learning rate as $5e-5$. We used AdamW optimizer and a linear learning rate schedule. We train the model for 10 epochs and report the result of the checkpoint with best validation accuracy, averaged across three random seeds.

Seq2Seq Models We use two variations (base, large) of T5 model (Raffel et al., 2020), which take the concatenation of question and answer sentences, and output the roles for each sentence sequentially. This model predicts the roles of all sentences in the answer as a single sequence. The input sequence is [$question$ [1] ans_1 [2] ans_2 ...], where ans_i denotes the i^{th} sentence in the answer, and the target output sequence is set to [[1] $role_1$ [2] $role_2$ [3]...], where $role_i$ is the corresponding role for ans_i (e.g. "Answer" for the Answer role). We limit the input/output to be 512/128 tokens. For evaluating the predicted roles, we parse the output string to identify the role predicted for each sentence. We used the batch size of 16, initial learning rate of $1e-4$ with AdamW optimizer and a linear learning rate schedule. We train the model for 30 epochs and report the result of the checkpoint with the best validation accuracy, averaged across three random seeds.

Human performance We provide two approximations for human performance: upperbound (u) and lowerbound (l). (1) Human (u): We compare each individual annotator's annotation with the majority label. This inflates human performance as one's own judgement affected the majority label. (2) Human (l): We compare all pairs of annotation and calculate average F1 and accuracy of all pairs. For **Match**, we compute the match for each annotation against the other two annotations.

6.3 Results

Table 4 reports the results on ELI5 test set.[11] All models outperform the majority and summary-lead baselines. The sequential prediction model

[11] Results on validation set are in Table 11 in the appendix.

(T5) significantly outperform classification model (RoBERTa) which makes a prediction per sentence. The roles with lower human agreement (auxiliary, organizational sentence, answer) also exhibit low model performances, reflecting the subjectivity and ambiguity of roles for some sentences. Overall, with a moderate amount of in-domain annotated data, our best model (T5-large) can reliably classify functional roles of sentences in the long-form answers, showing comparable performances to human lower bound.

Table 5 reports the results on the three out-of-domain datasets, WebGPT, NQ and ELI5-model (model-generated answers). Human agreement numbers are comparable across all datasets (0.53-0.59 for lower bound, 0.73-0.78 for upper bound). While T5-large still exhibits the best overall performance, all learned models perform worse, partially as the role distribution has changed. Despite trained on the ELI5 dataset, role classification model also perform worse on model-generated answers (ELI5-model), echoing our observation that human annotators find it challenging to process the discourse structure of model-generated answers. Our pilot showed that training with in-domain data improved the performances consistently, but the evaluation is on a small subset (after setting apart some for training), so we do not report it here. We anticipate that automatic role classification is feasible given moderate amount of annotation for all three human-written long-form answer datasets we study.

7 Related Work

Discourse structure. Our work is closely related to functional structures defined through content types explored in other domains; prior work has affirmed the usefulness of these structures in downstream NLP tasks. In news, Choubey et al. (2020) adopted Van Dijk (2013)'s content schema cataloging events (e.g., main event, anecdotal), which they showed to improve the performance of event coreference resolution. In scientific writing, content types (e.g., background, methodology) are shown to be useful for summarization (Teufel and Moens, 2002; Cohan et al., 2018), information extraction (Mizuta et al., 2006; Liakata et al., 2012), and information retrieval (Kircz, 1991; Liddy, 1991). The discourse structure of argumentative texts (e.g., support, rebuttal) (Peldszus and Stede, 2013; Becker et al., 2016; Stab and Gurevych, 2017) has also been applied on argumentation mining. To the best of our knowledge, no prior work has studied the discourse structure of long-form answers.

Question Answering. Recent work (Cao and Wang, 2021) have investigated the ontology of questions, which includes comparison questions, verification questions, judgement questions, etc. We construct the ontology of functional roles of answer sentences. One of the roles in our ontology is summary, yielding an extractive summarization dataset. This shares motivation with a line of work studying query-focused summarization (Xu and Lapata, 2020). Concurrent to our work, Su et al. (2022) studies improving faithfulness of long-form answer through predicting and focusing on salient information in retrieved evidence document. Lastly, our work build up on three datasets containing long-form answers (Kwiatkowski et al., 2019; Fan et al., 2019; Nakano et al., 2021) and extends the analysis of long-form answers from earlier studies (Krishna et al., 2021).

8 Conclusion

We present a linguistically motivated study of long-form answers. We find humans employ various strategies – introducing sentences laying out the structure of the answer, proposing hypothetical and real examples, and summarizing main points – to organize information. Our discourse analysis characterizes three types of long-form answers and reveals deficient discourse structures of model-generated answers. Discourse analysis can be fruitful direction for evaluating long-form answers. For instance, highlighting summary sentence(s) or sentence-level discourse role could be helpful for human evaluators to dissect long-form answers, whose length has been found to be challenging for human evaluation (Krishna et al., 2021). Trained role classifier can also evaluate the discourse structure of model-generated answers. Future work can explore using sentences belonging to the summary role to design evaluation metrics that focuses on the core parts of the answer (Nenkova and Passonneau, 2004), for assessing the *correctness* of generated the answer. Exploring controllable generation, such as encouraging models to provide summaries or examples, would be another exciting avenue for future work.

Ethical Considerations

We annotate existing, publicly available long-form question answering datasets which might contain incorrect and outdated information and societal biases. We collected annotations through crowdsourcing platform and also by recruiting undergraduate annotators at our educational institution. We paid a reasonable hourly wage ($13/hour) to annotators and documented our data collection process with datasheet (Gebru et al., 2021). We include studies on the extractiveness of long-form answers (how much content can be grounded to evidence document) through a coarse measure of lexical overlap. This is connected to faithfulness and reducing hallucination of QA system. Our study is limited to English sources, and we hope future work can address analysis in other languages.

Acknowledgements

This work was partially supported by NSF grants IIS-1850153, IIS-2107524. We thank Kalpesh Krishna and Mohit Iyyer for sharing the model predictions and human evaluation results. We would like to thank Tanya Goyal, Jiacheng Xu, Mohit Iyyer, anonymous reviewers and meta reviewer for providing constructive feedback to improve the draft. Lastly, we thank Maanasa V Darisi, Meona Khetrapal, Matthew Micyk, Misty Peng, Payton Wages, Sydney C Willett and crowd workers for their help with the complex data annotation.

References

Maria Becker, Alexis Palmer, and Anette Frank. 2016. Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 21–30, Berlin, Germany. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint 2002.08909*.

Jerry R Hobbs. 1985. On the coherence and structure of discourse.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, abs/2109.06835.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Joost G Kircz. 1991. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of documentation*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th international joint conference on natural language processing*, pages 605–613.

Andrew MacKinlay and Katja Markert. 2011. Modelling entity instantiations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 268–274.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *ArXiv*, abs/2004.03027.

A Appendix

A.1 Invalid QA

We provide definitions, as well as examples of each invalid QA type.

No valid answer The answer paragraph doesn't provide a valid answer to the question.

> **[Q]**: How does drinking alcohol affect your ability to lose weight?
> **[A]**: Alcohol itself is extremely calorically dense.Doesn't really matter whether you're drinking a light beer or shots, alcohol itself has plenty of calories. Just think of every three shots as eating a mcdouble, with even less nutritional value.

Nonsensical question The question is nonsensical and it is unclear what is asked.

> **[Q]**: asia vs rest of the world cricket match

Multiple questions asked More than one question are asked in the question sentence.

> **[Q]**: what is a limpet and where does it live

Assumptions in the question rejected The answer focuses on rejecting assumptions in the question, without answering the question.

> **[Q]**: Why is it that as we get older, we are able to handle eating hotter foods
> **[A]**: I'm not sure I accept the premise.Children in cultures where spicy food is common, think nothing of it.My nephews had no problem eating hot peppers when they were very young because it was just a normal part of their diet.[...]

A.2 Role annotation

We include example role annotations in Table 6 which demonstrate disagreement between Auxiliary Information and the Answer role. Sentence 2 in answer (a) was annotated as answer by most of the annotators as it elaborates on becoming a legal 'next of kin' by providing a counterfactual scenario. One annotator annotated it as auxiliary as it touches upon how the decisions would be up to the parents, which goes beyond what is asked in the question. For answer (b), while most annotators think that sentence 1 is of Answer role, one annotator annotated it as Auxiliary Information which only talks about the property of purple.

A.3 Implementation Details

We use `pytorch-transformers` Wolf et al. (2019) to implement our models. The hyperparameters are manually searched by the authors.

idx	(a) Question: What are the benefits of marriage in the U.S.?	Role	Other annotation
1	I think one of the biggest ones is that your spouse becomes your legal 'next of kin', meaning you can make medical decisions for them, own their property after they die, etc.	Summary	
2	If you aren't married you are not legally a part of that person's life, so any legal or medical decisions would be up to the parents of that individual.	Answer	Auxiliary
3	That's why marriage equality was important a few years ago.	Auxiliary	
4	If someone was with their partner for 15 years and then suddenly dropped dead, their partner had better hope their in-laws liked them or even supported the partnership in the first place.	Example	Auxiliary
5	If not, the parents could just take the house and all the money (provided the person didn't have a will).	Example	Auxiliary
6	There are probably other benefits, but I think this is one of the big ones..	Answer	Misc

idx	(b) Question: what is the difference of purple and violet	Role	Other annotation
1	Purple is a color intermediate between blue and red .	Answer	Auxiliary
2	It is similar to violet , but unlike violet , which is a spectral color with its own wavelength on the visible spectrum of light , purple is a composite color made by combining red and blue .	Summary	
3	According to surveys in Europe and the U.S. , purple is the color most often associated with royalty , magic , mystery , and piety .	Auxiliary	
4	When combined with pink , it is associated with eroticism , femininity , and seduction .	Auxiliary	

Table 6: Question paired with their paragraph level answer. Each sentence in a paragraph level answer is annotated with its role defined in Section 2. We also include the other annotated role to demonstrate cases where annotators disagree with each other. (a) is from ELI5 dataset, (b) is from NQ dataset.

when did the temperance movement begin in the united states
what are the ingredients in chili con carne
is pink rock salt the same as sea salt
why is muharram the first month of the islamic calendar
what qualifies a citizen in the han dynasty to hold a government job
what is the difference between cheddar and american cheese

Table 7: Examples of NQ long questions classified as factoid (top) v.s. non-factoid (bottom).

Role	1-gram		2-gram		# Sentences	
	E	W	E	W	E	W
Sum	**0.10**	0.65	0.01	0.36	547	192
Ans	**0.10**	0.61	0.01	0.32	571	154
Aux	0.08	**0.68**	0.00	**0.41**	319	146
Ex	0.07	0.65	0.00	0.39	262	43
Misc	0.03	-	0.00	-	199	-
Org	0.07	0.54	0.01	0.29	19	16

Table 8: Unigram and bigram overlap between the answer sentence and a paired evidence for **E**LI5 and **W**ebGPT per role. The last column shows the number of annotated sentences belonging to the specific role.

Question classification model A difficulty in repurposing NQ is that not all questions with paragraph answers *only* actually need multiple sentences. To identify complex questions, we built a simple BERT-based classifier, trained to distinguish NQ questions with short answers (i.e., less than five tokens) and ELI5 questions. We use the [CLS] token from BERT model to perform prediction. We use the original split from the ELI5 dataset, and split the NQ open's validation set into val and test set. We preprocessed the questions by converting to lowercase and exclude punctuation to remove syntactic differences between ELI5 and NQ questions. We fine-tuned the `bert-base-uncased` model for 3 epochs, with an initial learning rate of $5e-5$ and batch size of 32. We use the model with the highest validation F1 as the question classifier, which achieves F1 of 0.97 and 0.94 on validation and test set respectively. We then run this classifier to select the non factoid questions from NQ questions with long-form answers, which classifies around 10%, out of the 27,752 NQ long questions as non-factoid. Examples are in Table 7.

A.4 Annotation Interface

Figure 5, 6, 7, 8 9, and 10 show the annotation guideline as well as interface presented to the annotators (we present Step 1 for crowdworkers, Step 2 and Step 3 for student annotators). We didn't capture the extended example section as well as FAQ here due to space.

Reason	% answer	Fleiss Kappa	Pairwise Agreement
ELI5-model answers			
No valid answer	39%	0.55	0.82
Nonsensical question	1%	0	0.99
Multiple questions	6%	0.33	0.96
Rejected presupposition	8%	0.28	0.95
ELI5, WebGPT and NQ answers			
No valid answer	11%	0.60	0.99
Nonsensical question	0%	0.67	0.99
Multiple questions	5%	0.78	0.99
Rejected presupposition	8%	0.33	0.99

Table 9: Different reasons for invalid question answer pairs for ELI5-model and annotator agreement. We report both Fleiss kappa and pairwise agreement after reannotation. For reference, we also report agreement for human-written answers annotated.

Question: Do animals know they're going to die?	**Role**
I read an article about this once, I can't find it now, but I remember reading about a dog that had been put into a room with a vacuum cleaner, and it didn't notice it was sucking in air, it just started sucking in air as normal.	Example
It was pretty amazing to watch.	*Disagreed*
So it was just sucking in air.	Example
Then, the dog got out of the room and began running around the house, running into things and being hurt.	Example
It eventually just died of exhaustion.	Example
So, no, they don't know.	Answer
But it is interesting to think about.	Miscellaneous
It might have just been a part of their routine, or it might have been a learned behavior, or it might have been something they did because it was the only way they could do it, and they figured it out, and it was just a part of their routine, and they thought it was cool.	Answer

Table 10: An example of model-generated answer with sentence-level role annotation.

System	Acc	Match	Ma-F1	Ans	Sum	Aux	Ex	Org	Msc
Majority	0.29	0.44	0.07	0	0.44	0	0	0	0
Summary	0.34	0.57	0.14	0.43	0.43	0	0	0	0
RoBERTa	0.45	0.63	0.52	0.38	0.57	0.31	0.54	0.67	0.56
T5-base	0.53	0.74	0.56	0.49	0.54	0.33	0.64	0.56	0.76
T5-large	**0.57**	**0.78**	**0.64**	**0.50**	**0.58**	**0.46**	**0.71**	**0.89**	**0.71**
Human (l)	0.57	0.73	0.57	0.50	0.66	0.38	0.68	0.47	0.73
Human (u)	0.76	1.00	0.65	0.72	0.82	0.67	0.83	0.69	0.85

Table 11: Role identification results on validation split of ELI5 dataset.

System	Ans	Sum	Aux	Ex	Org	Msc
Majority	0/0/0	0.52/0.52/0.36	0/0/0	0/0/0	0/0/0	-/0/0
Summary	0.45/0.35/0.44	0.4/0.53/0.51	0/0/0	0/0/0	0/0/0	-/0/0
RoBERTa	0.40/0.19/0.42	0.48/0.55/0.52	0.20/0.46/0.43	0.46/0.41/0.49	0.08/0.00/0.17	-/0.38/0.31
T5-base	0.47/0.33/0.46	0.45/0.52/0.48	0.26/0.48/0.27	**0.55**/0.31/0.48	0.14/0.00/0.37	-/0.44/0.4
T5-large	**0.49/0.32/0.47**	**0.51/0.54/0.53**	**0.38/0.49**/0.36	0.51/**0.40/0.57**	**0.43/0.06/0.49**	-/**0.58/0.43**
Human (l)	0.40/0.35/0.49	0.62/0.70/0.61	0.54/0.65/0.57	0.47/0.49/0.71	0.65/0.10/0.60	-/0.27/0.50
Human (u)	0.66/0.63/0.75	0.79/0.85/0.81	0.74/0.82/0.79	0.69/0.71/0.85	0.80/0.41/0.78	-/0.54/0.72

Table 12: Per role performance on three out-of-domain datasets. The three numbers in each cell represents performance on WebGPT, NQ, ELI5-model in order.

Understanding the structure of answer for complex queries

Thank you for participating in this task! We would like to better understand what composes an answer to complex queries such as "Why do birds sing in the morning?". You will be presented with a question, originally posted on Reddit forum or entered in Google, and an answer paragraph, written by another user or selected from a Wikipedia page. Your job is to provide a label for each sentence in the answer paragraph, deciding the role of this sentence.

There are three steps in this task:

Step 1: You will determine if this sample is a valid (question, answer) pair based on a set of reasons we defined below. If it is invalid, please simply choose a reason and skip Step 2.

Step 2: You will label each sentence in the answer paragraph with a role. You will choose from a set of roles we defined, which will be explained in the instruction.

Step 3: You will choose one or more sentences which contain the main answer to the question.

Figure 5: Screenshot of annotation guideline (overall).

Step 1: Validity Check

In the first step, you will determine if this is a valid (question, answer) pair. We define below reasons for which a (question, answer) pair should be considered invalid for this task:

Reason	Defintion	Example
No valid answer	The answer paragraph doesn't provide a valid answer to the question. Please note that if the answer can be inferred from the paragraph, it should be a valid answer (see second example).	*(1) Example of no valid answer* **Question:**Why can't we tickle ourselves? **Answer:**I can. Am I a mutant? Is this my power? **Explanation:** There is no valid answer about the question asked. *(2) Counterexample: Implicit answer* **Question:**what is the difference between the qx56 and qx80 **Answer:**The Infiniti QX80 (called the Infiniti QX56 until 2014) is a full - size luxury SUV built by Nissan Motor Company Infiniti division . The naming convention originally adhered to the current trend of using a numeric designation derived from the engine 's displacement , thus QX56 since the car has a 5.6 - liter engine . From the 2014 model year , the car was renamed the QX80 , as part of Infiniti 's model name rebranding .The new name carries no meaning beyond suggesting that the vehicle is larger than smaller models such as the QX60 . **Explanation:**The answer could be inferred from the first sentence (QX80 is a new name of QX56).
Nonsensical question	The question is nonsensical and it is unclear what is asked. Please note that some queries don't contain any question word, and yet it is clear what they are asking about (see second example) and hence they shouldn't fall under this category.	*(1) Nonsensical question* **Question:**i'm glad i'm not a kennedy **Explanation:**This is not a question. *(2) Counterexample: Short queries* **Question:**difference between senate and house of commons canada **Explanation:**This query doesn't contain any question word, but it if clear that it is asking about "What is the difference between senate and house of commons in Canada".
Multiple questions asked	More than one questions are asked in the answer question. Please note that sometimes the question can be complicated and yet **only one question is asked** (See second example) and hence they shouldn't fall under this category.	*(1) Multiple questions asked* **Question:**How light pollution works, and would the stars appear straight away if we turned the lights off on the entire globe? **Explanation:** Two questions are asked in this sentence. We exclude such questions since it will be difficult to select the main content. *(2) Counterexample: One complicated question* **Question:**Is the presence of pain evidence that damage is done, or is it just a warning? **Explanation:**This is *one* disjunctive question, asking whether the presence of pain is an evidence that damage is done or a warning.
Assumptions in the question rejected	The answer focuses on rejecting assumptions in the question, without answering the question. We would like to point out two counter examples (1)If it is a question asking about difference between two or more entities, an answer suggesting that there is no difference should be considered as a **valid QA**, instead of "Assumptions in the question rejected" (see second example). (2)If the answer *corrects* some assumption from the question, but still provides answer of what is asked (see third example).	*(1) Assumption rejected:* **Question:**Why can't I share a puddle of water with my dog? I get sick every time but he's fine. **Answer:**dirty water' is pretty vague. There are lots of different diseases and contamits that water can have. A singular adaptation is not likely to make you 'dirty water' proof. Your dog himself is not dirty water proof. He can potentially get sickened or parasitized by contaminated water as well. **Explanation:** The answer argues that dogs can also be sickened by contaminated water which rejects the assumption in the question "I get sick every time but he's fine". *(2) Counterexample: question about difference* **Question:**what's the difference between 9mm luger and parabellum **Answer:**The 9 × 19mm Parabellum is a firearms cartridge that was designed by Georg Luger [...] it is designated as the 9mm Luger by the SAAMI , and the 9 mm Luger by the C.I.P. . **Explanation:** The answer suggests that there is no difference between 9mm luger and parabellum. In this case, we treat this as a **valid** QA pair. *(3) Counterexample: Partial rejection* **Question:**How can the chlorine in tap water kill microorganisms, but not the cells in our bodies? **Answer:**It will kill cells in the body at high enough concentrations. Scientist and experts do their best to use as little as possible to get the job done and use "just enough" to get rid of the bacteria. Also, chlorine evaporates and degrades fairly quickly. So from the time the water is intitially treated to the time it eventually gets to your taps a lot of the chlorine is gone. **Explanation:** The answer corrects the assumption that chlorine can kill cells in the body if the concentration is high. However, it also explains why the **chlorine in tap water** doesn't, which is the answer to the question.

Figure 6: Screenshot of annotation guideline (Step 1).

Step 2: Assigning Roles

In the second step, you will assign one of the roles defined below to each sentence in the answer paragraph.

Role	Definition	Example
Answer	The sentence contains information asked in the question.	Q: what were the significant outcomes of the war of 1812 on the united states A: [...]**The main result of the war was two centuries of peace between the United States and Britain.**[...]
Answer - Example	The sentence provides a specific example (either a concrete one or hypothetical one) for the answer. Note that sometimes the sentence will contain indication for example, (e.g. "For example, ..."), sometimes there might not be such a clear indication. For sentence that doesn't explicitly indicate an example (see Example 1), please select this role only if inserting "for example" to the sentence still makes it a coherent sentence in the paragraph.	*(1) Hypothetical example* Q: Were major news outlets established with political bias or was it formed over time? A:[...]This is impossible due to the problem of "anchoring."**Consider a world where people on the right want the tax rate to be 1% lower and people on the left want the tax rate to be 1% higher.**.[...] *(2)Concrete example* Q: What is it about electricity that kills you? A:[...]**For example, static electricity consists of tens of thousands of volts, but basically no amps.** [...]
Answer - Organizational sentence	The sentence either tells the readers what they should expect to read in the answer, or indicates that the answer will come in the following sentences.	*(1) What should expect in the answer* Q: Why do candidates like O'Malley and Kasich stay in the race when polls consistently show them at around 2%? A:**There are a few reasons candidates with "no chance" to win keep running.** [...] *(2) Indication of answer in the following sentences* Q: Why do i sometimes get what seems to be a mosquito bite on my face at night in the middle of winter? A:It might actually be a mosquito bite. I find the odd mosquito in my house in the winter from time to time, and I'm in Canada.[...]**So why does it happen more often when you shower?**It's largely because the pressure from the showerhead, plus the fact that if you're shampooing or conditioning you're going to be running your fingers through it a lot, means that any loose hairs -- such as those in the telogen phase -- are more likely to come out, and in greater numbers.
Auxiliary Information	The sentence provides information that is relevant to what is discussed in the answer, but not explicitly asked in the question. Such information might be (1) background information that is helpful for readers who are not familiar with the topic discussed (2) related information that extends on what is asked in the question. Check out FAQ for more examples and distinction with role "Answer".	*(1)Background information* Q: Why is it better to use cloning software instead of just copying and pasting the entire drive? A: **When you install an operating system, it sets up what's called a master file table, which holds file permissions for your files (telling the computer who can access them) among other things, which are important for the OS to work properly.** [...] Simply copy-pasting files doesn't copy either of these, meaning if you want to back up an OS installation you should clone the disk instead. *(2)Related information* Q: what is the difference between mandi and kabsa A: Meat for kabsa can be cooked in various ways .A popular way of preparing meat is called mandi .**This is an ancient technique that originates in Yemen , whereby meat is barbecued in a deep hole in the ground that is covered while the meat cooks . Another way of preparing and serving meat for kabsa is mathbi , where seasoned meat is grilled on flat stones that are placed on top of burning embers .**
Miscellaneous	The sentence serves other roles than the ones mentioned above. Here we list several scenarios that we have observed in the example. Note that this is not the complete list and it is possible that a sentence serves other roles than those listed here.	*(1) Stating limitation of the answer* Q: How do movies become Oscar contenders? A:**The academy has some pretty extensive rules, but heres the short eli5 version 1.**[...] *(2) Providing sources of the answer* Q: Why do birds chirp more in the morning? A:**This info comes from the breeder who sold me my lovebird, so I can't really provide a source.**[...] *(3) Expressing sentiments* Q: Why did Catholicism embrace celibacy and why did Protestantism reject it? A:**Good God, the amount of misinformation upvoted is hurting.** [...] *(4) Providing pointers to other resources* Q: Why did Catholicism embrace celibacy and why did Protestantism reject it? A: **/r/askhistorians has a few excellent discussions about this.** [...]

Figure 7: Screenshot of annotation guideline (Step 2).

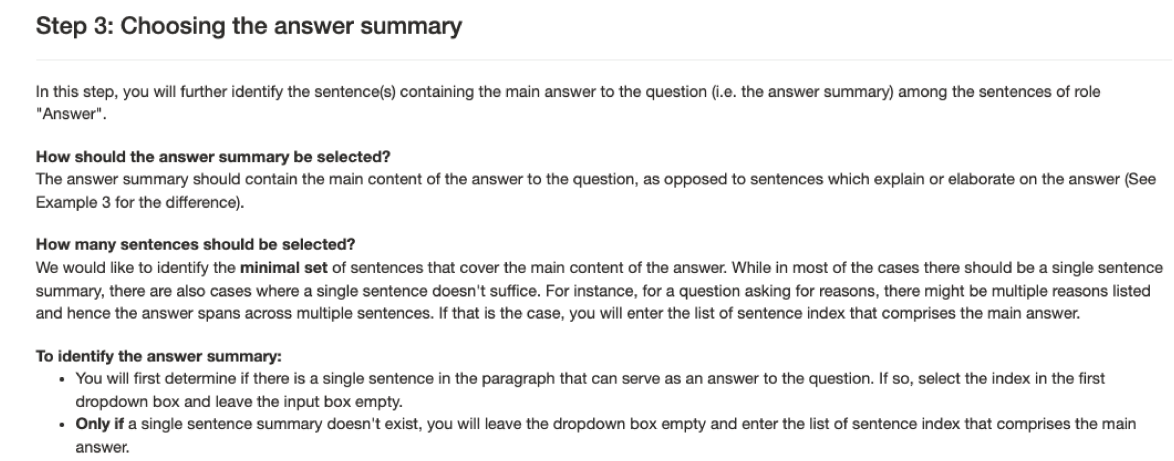

Figure 8: Screenshot of annotation guideline (Step 3).

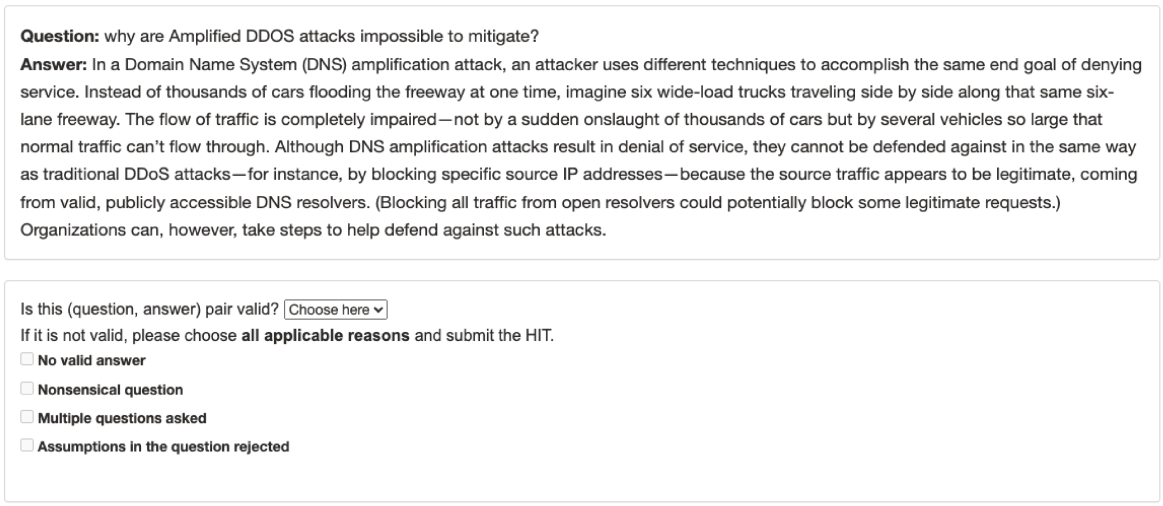

Figure 9: Screenshot of annotation interface for question validity.

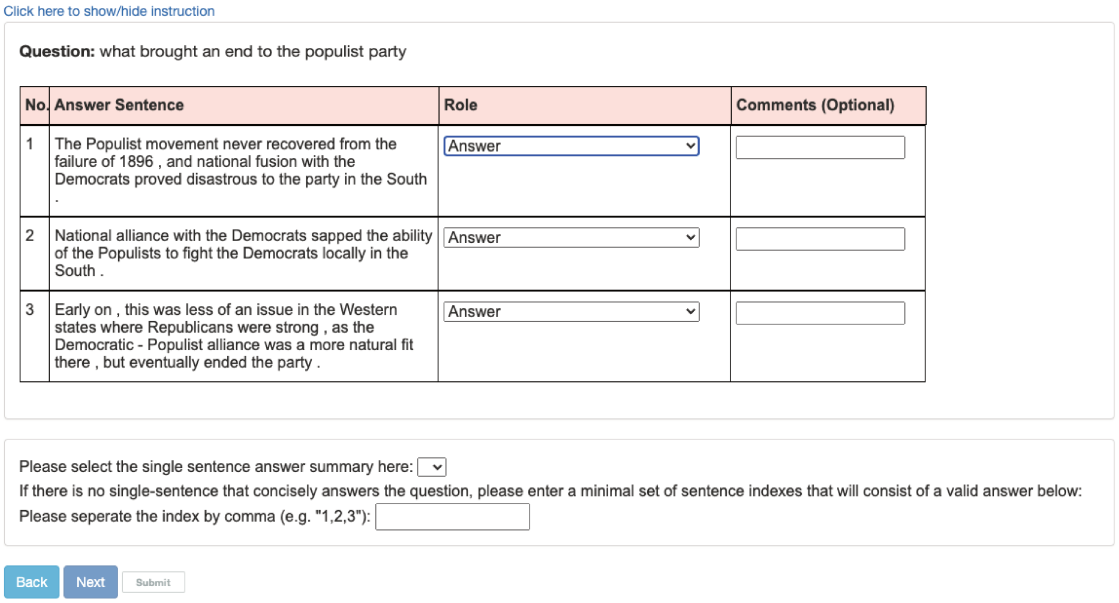

Figure 10: Screenshot of annotation interface for sentence-level role, as well as summary sentence selection.