

# Finding Generalizable Evidence by Learning to Convince Q&A Models

Ethan Perez<sup>†</sup> Siddharth Karamcheti<sup>‡</sup>  
Rob Fergus<sup>††</sup> Jason Weston<sup>††</sup> Douwe Kiela<sup>‡</sup> Kyunghyun Cho<sup>††\*</sup>  
<sup>†</sup>New York University, <sup>‡</sup>Facebook AI Research, <sup>\*</sup>CIFAR Azrieli Global Scholar  
perez@nyu.edu

## Abstract

We propose a system that finds the strongest supporting evidence for a given answer to a question, using passage-based question-answering (QA) as a testbed. We train evidence agents to select the passage sentences that most convince a pretrained QA model of a given answer, if the QA model received those sentences instead of the full passage. Rather than finding evidence that convinces one model alone, we find that agents select evidence that generalizes; agent-chosen evidence increases the plausibility of the supported answer, as judged by other QA models and humans. Given its general nature, this approach improves QA in a robust manner: using agent-selected evidence (i) humans can correctly answer questions with only  $\sim 20\%$  of the full passage and (ii) QA models can generalize to longer passages and harder questions.

## 1 Introduction

There is great value in understanding the fundamental nature of a question (Chalmers, 2015). Distilling the core of an issue, however, is time-consuming. Finding the correct answer to a given question may require reading large volumes of text or understanding complex arguments. Here, we examine if we can automatically discover the underlying properties of problems such as question answering by examining how machine learning models learn to solve that task.

We examine this question in the context of passage-based question-answering (QA). Inspired by work in interpreting neural networks (Lei et al., 2016), we have agents find a subset of the passage (i.e., supporting evidence) that maximizes a QA model’s probability of a particular answer. Each agent (one agent per answer) finds the sentences that a QA model regards as strong evidence for its answer, using either exhaustive search or learned prediction. Figure 1 shows an example.

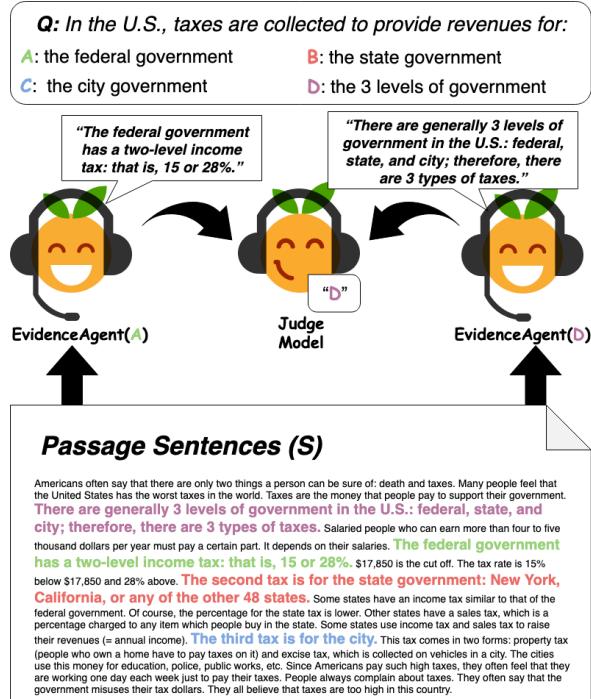


Figure 1: Evidence agents quote sentences from the passage to convince a question-answering judge model of an answer.

To examine to what extent evidence is general and independent of the model, we evaluate if humans and other models find selected evidence to be valid support for an answer too. We find that, when provided with evidence selected by a given agent, both humans and models favor that agent’s answer over other answers. When human evaluators read an agent’s selected evidence in lieu of the full passage, humans tend to select the agent-supported answer.

Given that this approach appears to capture some general, underlying properties of the problem, we examine if evidence agents can be used to assist human QA and to improve generalization of other QA models. We find that humans can accurately answer questions on QA benchmarks, based

on evidence for each possible answer, using only 20% of the sentences in the full passage. We observe a similar trend with QA models: using only selected evidence, QA models trained on short passages can generalize more accurately to questions about longer passages, compared to when the models use the full passage. Furthermore, QA models trained on middle-school reading comprehension questions generalize better to high-school exam questions by answering only based on the most convincing evidence instead of the full passage. Overall, our results suggest that learning to select supporting evidence by having agents try to convince a judge model of their designated answer improves QA in a general and robust way.

## 2 Learning to Convince Q&A Models

Figure 1 shows an overview of the problem setup. We aim to find the passage sentences that provide the most convincing evidence for each answer option, with respect to a given QA model (the *judge*). To do so, we are given a sequence of passage sentences  $S = [S(1), \dots, S(m)]$ , a question  $Q$ , and a sequence of answer options  $A = [A(1), \dots, A(n)]$ . We train a judge model with parameters  $\phi$  to predict the correct answer index  $i^*$  by maximizing  $p_\phi(\text{answer} = i^* | S, Q, A)$ .

Next, we assign each answer  $A(i)$  to one evidence agent,  $\text{AGENT}(i)$ .  $\text{AGENT}(i)$  aims to find evidence  $E(i)$ , a subsequence of passage sentences  $S$  that the judge finds to support  $A(i)$ . For ease of notation, we use set notation to describe  $E(i)$  and  $S$ , though we emphasize these are ordered sequences.  $\text{AGENT}(i)$  aims to maximize the judge's probability on  $A(i)$  when conditioned on  $E(i)$  instead of  $S$ , i.e.,  $\arg\max_{E(i) \subseteq S} p_\phi(i | E(i), Q, A)$ . We now describe three different settings of having agents select evidence, which we use in different experimental sections (§4-6).

**Individual Sequential Decision-Making** Since computing the optimal  $E(i)$  directly is intractable, a single  $\text{AGENT}(i)$  can instead find a reasonable  $E(i)$  by making  $T$  sequential, greedy choices about which sentence to add to  $E(i)$ . In this setting, the agent ignores the actions of the other agents. At time  $t$ ,  $\text{AGENT}(i)$  chooses index  $e_{i,t}$  of the sentence in  $S$  such that:

$$e_{i,t} = \arg\max_{1 \leq e' \leq |S|} p_\phi(i | \{S(e')\} \cup E(i, t-1), Q, A), \quad (1)$$

where  $E(i, t)$  is the subsequence of sentences in  $S$  that  $\text{AGENT}(i)$  has chosen until time step  $t$ , i.e.,  $E(i, t) = \{S(e_{i,t})\} \cup E(i, t-1)$  with  $E(i, 0) = \emptyset$  and  $E(i) = E(i, T)$ . It is a no-op to add a sentence  $S(e_{i,t})$  that is already in the selected evidence  $E(i, t-1)$ . The individual decision-making setting is useful for selecting evidence to support one particular answer.

**Competing Agents: Free-for-All** Alternatively, multiple evidence agents can compete at once to support unique answers, by each contributing part of the judge's total evidence. Agent competition is useful as agents collectively select a pool of question-relevant evidence that may serve as a summary to answer the question. Here, each of  $\text{AGENT}(1), \dots, \text{AGENT}(n)$  finds evidence that would convince the judge to select its respective answer,  $A(1), \dots, A(n)$ .  $\text{AGENT}(i)$  chooses a sentence  $S(e_{i,t})$  by conditioning on all agents' prior choices:

$$e_{i,t} = \arg\max_{1 \leq e' \leq |S|} p_\phi(i | \{S(e')\} \cup E(*, t-1), Q, A),$$

需要考虑其他 AGENT

where  $E(*, t-1) = \bigcup_{j=1}^n E(j, t-1)$ . 在 t 时刻之前选择的 evidence sentence.

Agents simultaneously select a sentence each, doing so sequentially for  $t$  time steps, to jointly compose the final pool of evidence. We allow an agent to select a sentence previously chosen by another agent, but we do not keep duplicates in the pool of evidence. Conditioning on other agents' choices is a form of interaction that may enable competing agents to produce a more informative total pool of evidence. More informative evidence may enable a judge to answer questions more accurately without the full passage.

**Competing Agents: Round Robin** Lastly, agents can compete round robin style, in which case we aggregate the outcomes of all  $\binom{n}{2}$  pairs of answers  $\{A(i), A(j)\}$  competing. Any given  $\text{AGENT}(i)$  participates in  $n - 1$  rounds, each time contributing half of the sentences given to the judge. In each one-on-one round, two agents select a sentence each at once. They do so iteratively multiple times, as in the free-for-all setup. To aggregate pairwise outcomes and compute an answer  $i$ 's probability, we average its probability over all rounds involving  $\text{AGENT}(i)$ :

$$\frac{1}{n-1} \sum_{j=1}^n \mathbb{1}(i \neq j) * p_\phi(i | E(i) \cup E(j), Q, A)$$

## 2.1 Judge Models

The judge model is trained on QA, and it is the model that the evidence agents need to convince. We aim to select diverse model classes, in order to: (i) test the generality of the evidence produced by learning to convince different models; and (ii) to have a broad suite of models to evaluate the agent-chosen evidence. Each model class assigns every answer  $A(i)$  a score, where the predicted answer is the one with the highest score. We use this score  $L(i)$  as a softmax logit to produce answer probabilities. Each model class computes  $L(i)$  in a different manner. In what follows, we describe the various judge models we examine.

**TFIDF** We define a function  $\text{BoW}_{\text{TFIDF}}$  that embeds text into its corresponding TFIDF-weighted bag-of-words vector. We compute the cosine similarity of the embeddings for two texts  $\mathbf{X}$  and  $\mathbf{Y}$ :

$$\text{TFIDF}(\mathbf{X}, \mathbf{Y}) = \cos(\text{BoW}_{\text{TFIDF}}(\mathbf{X}), \text{BoW}_{\text{TFIDF}}(\mathbf{Y}))$$

We define two model classes that select the answer most similar to the input passage sentences:  $L(i) = \text{TFIDF}(\mathcal{S}, [Q; A(i)])$ , and  $L(i) = \text{TFIDF}(\mathcal{S}, A(i))$ . passage 通过直接选择

**fastText** We define a function  $\text{BoW}_{\text{FT}}$  that computes the average bag-of-words representation of some text using *fastText* embeddings (Joulin et al., 2017). We use 300-dimensional *fastText* word vectors pretrained on Common Crawl. We compute the cosine similarity between the embeddings for two texts  $\mathbf{X}$  and  $\mathbf{Y}$  using:

$$\text{fastText}(\mathbf{X}, \mathbf{Y}) = \cos(\text{BoW}_{\text{FT}}(\mathbf{X}), \text{BoW}_{\text{FT}}(\mathbf{Y}))$$

This method has proven to be a strong baseline for evaluating the similarity between two texts (Perone et al., 2018). Using this function, we define a model class that selects the answer most similar to the input passage context:  $L(i) = \text{fastText}(\mathcal{S}, A(i))$ .

**BERT**  $L(i)$  is computed using the multiple-choice adaptation of BERT (Devlin et al., 2019; Radford et al., 2018; Si, 2019), a pre-trained transformer network (Vaswani et al., 2017). We fine-tune all BERT parameters during training. This model predicts  $L(i)$  using a trainable vector  $v$  and BERT’s first token embedding:  $L(i) = v^\top \text{BERT}([\mathcal{S}; Q; A(i)])$ . 取[CLS]处的向量

We experiment with both the BERT<sub>BASE</sub> model (12 layers) and BERT<sub>LARGE</sub> (24 layers). For training details, see Appendix B.

## Predicting | Loss Target

	<i>CE</i>	$S(e_{i,t})$
Search	<i>MSE</i>	$p_\phi(i S(e')) \cup E(i, t), Q, A)$
$p(i)$	<i>MSE</i>	$p_\phi(i S(e')) \cup E(i, t), Q, A)$ $- p_\phi(i E(i, t), Q, A)$
$\Delta p(i)$		

Table 1: The loss functions and prediction targets for three learned agents. *CE*: cross entropy. *MSE*: mean squared error.  $e'$  takes on integer values from 1 to  $|S|$ .

## 2.2 Evidence Agents

In this section, we describe the specific models we use as evidence agents. The agents select sentences according to Equation 1, either exactly or via function approximation.

**Search agent**  $\text{AGENT}(i)$  at time  $t$  chooses the sentence  $S(e_{i,t})$  that maximizes  $p_\phi(i|S(i, t), Q, A)$ , after exhaustively trying each possible  $S(e_{i,t}) \in S$ . Search agents that query TFIDF or fastText models maximize TFIDF or fastText scores directly (i.e.,  $L(i)$ , rather than  $p_\phi(i|S(i, t), Q, A)$ ).

**Learned agent** We train a model to predict how a sentence would influence the judge’s answer, instead of directly evaluating answer probabilities at test time. This approach may be less prone to selecting sentences that exploit hard-to-predict quirks in the judge; humans may be less likely to find such sentences to be valid evidence for an answer (discussed in §4.1). We define several loss functions and prediction targets, shown in Table 1. Each forward pass, agents predict one scalar per passage sentence via end-of-sentence token positions. We optimize these predictions using Adam (Kingma and Ba, 2015) on one loss from Table 1. For  $t > 1$ , we find it effective to simply predict the judge model at  $t = 1$  and use this distribution for all time steps during inference. This trick speeds up training by enabling us to precompute prediction targets using the judge model, instead of querying it constantly during training.

We use BERT<sub>BASE</sub> for all learned agents. Learned agents predict the BERT<sub>BASE</sub> judge, as it is more efficient to compute than BERT<sub>LARGE</sub>. Each agent  $\text{AGENT}(i)$  is assigned the answer  $A(i)$  that it should support. We train one learned agent to find evidence for an arbitrary answer  $i$ . We condition  $\text{AGENT}(i)$  on  $i$  using a binary indicator when predicting  $L(i)$ . We add the indicator to BERT’s first token segment indicator and embed it

into vectors  $\gamma$  and  $\beta$ ; for each timestep’s features  $f$  from BERT, we scale and shift  $f$  element-wise:  $(\gamma * f) + \beta$  (Perez et al., 2018; Dumoulin et al., 2018). See Appendix B for training details.

Notably, learning to convince a judge model does not require answer labels to a question. Even if the judge only learns from a few labeled examples, evidence agents can learn to model the judge’s behavior on more data and out-of-distribution data without labels.

### 3 Experimental Setup

#### 3.1 Evaluating Evidence Agents

**Evaluation Desiderata** An ideal evidence agent should be able to find evidence for its answer w.r.t. a judge, regardless (to some extent) of the specific answer it defends. To appropriately evaluate evidence agents, we need to use questions with more than one defensible, passage-supported answer per question. In this way, an agent’s performance will not depend disproportionately on the answer it is to defend, rather than its ability to find evidence.

**Multiple-choice QA: RACE and DREAM** For our experiments, we use RACE (Lai et al., 2017) and DREAM (Sun et al., 2019), two multiple-choice, passage-based QA datasets. Both consist of reading comprehension exams for Chinese students learning English; teachers explicitly designed answer options to be plausible (even if incorrect), in order to test language understanding. Each question has 4 total answer options in RACE and 3 in DREAM. Exactly one option is correct. DREAM consists of 10K informal, dialogue-based passages. RACE consists of 100K formal, written passages (i.e., news, fiction, or well-written articles). RACE also divides into easier, middle school questions (29%) and harder, high school questions (71%).

**Other datasets we considered** Multiple-choice passage-based QA tasks are well-suited for our purposes. Multiple-choice QA allows agents to support clear, dataset-curated possible answers. In contrast, Sugawara et al. (2018) show that 5–20% of questions in extractive, span-based QA datasets have only one valid candidate option. For example, some “when” questions are about passages with only one date. Sugawara et al. argue that multiple-choice datasets such as RACE do not have this issue, as answer candidates are manually created. In preliminary experiments on

Judge Model	RACE	DREAM
Random	25.0	33.3
TFIDF( $S, [Q; A]$ )	32.6	44.4
TFIDF( $S, A$ )	31.6	44.5
fastText( $S, A$ )	30.4	38.4
BERT <sub>BASE</sub>	65.4	61.0
BERT <sub>LARGE</sub>	69.4	64.9
Human Adult*	94.5	98.6

Table 2: RACE and DREAM test accuracy of various judge models using the full passage. Our agents use these models to find evidence. The models cover a spectrum of QA ability. (\*) reports ceiling accuracy from original dataset papers.

SQuAD (Rajpurkar et al., 2016), we found that agents could only learn to convince the judge model when supporting the correct answer (one answer per question).

#### 3.2 Training and Evaluating Models

Our setup is not directly comparable to standard QA setups, as we aim to evaluate evidence rather than raw QA accuracy. However, each judge model’s accuracy is useful to know for analysis purposes. Table 2 shows model accuracies, which cover a broad range. BERT models significantly outperform word-based baselines (TFIDF and fastText), and BERT<sub>LARGE</sub> achieves the best overall accuracy. No model achieves the estimated human ceiling for either RACE (Lai et al., 2017) or DREAM (Sun et al., 2019).

Our code is available at <https://github.com/ethanjperez/convince>. We build off AllenNLP (Gardner et al., 2018) using PyTorch (Paszke et al., 2017). For all human evaluations, we use Amazon Mechanical Turk via ParlAI (Miller et al., 2017). Appendix B describes preprocessing and training details.

### 4 Agents Select General Evidence

#### 4.1 Human Evaluation of Evidence

Would evidence that convinces a model also be valid evidence to humans? On one hand, there is ample work suggesting that neural networks can learn similar patterns as humans do. Convolutional networks trained on ImageNet share similarities with the human visual cortex (Cadieu et al., 2014). In machine translation, attention learns to align foreign words with their native counterparts (Bahdanau et al., 2015). On the other hand, neural networks often do not behave as humans

Evidence Sentence Selection Method		How Often Human Selects Agent’s Answer (%)					
		RACE			DREAM		
		Overall	Agent Answer is Right	Agent Answer is Wrong	Overall	Agent Answer is Right	Agent Answer is Wrong
<b>Baselines</b>	No Sentence Given	25.0	52.5	15.8	33.3	43.3	28.4
	Human Selection	41.6	75.1	30.4	50.7	84.9	33.5
<b>Search Agents querying...</b>	TFIDF( $S, [Q; A(i)]$ )	33.5	69.6	21.5	41.7	68.8	28.1
	fastText( $S, A(i)$ )	37.1	74.2	24.7	41.5	75.6	24.5
	TFIDF( $S, A(i)$ )	38.0	71.4	26.9	43.4	75.2	27.6
	BERT <sub>BASE</sub>	38.4	68.4	28.4	50.5	82.5	34.6
	BERT <sub>LARGE</sub>	40.1	71.0	29.9	52.3	79.4	38.7
<b>Learned Agents predicting...</b>	Search	40.0	71.0	29.7	49.1	78.3	34.6
	$p(i)$	42.0	74.6	31.1	50.0	77.3	36.3
	$\Delta p(i)$	41.1	73.2	30.4	48.2	76.5	34.0

Table 3: *Human evaluation*: **Search Agents** select evidence by querying the specified judge model, and **Learned Agents** predict the strongest evidence w.r.t. a judge model ( $\text{BERT}_{\text{BASE}}$ ); humans then answer the question using the selected evidence sentence (without the full passage). Most agents do on average find evidence for their answer, right or wrong. Agents are more effective at supporting right answers.

do. Neural networks are susceptible to adversarial examples—changes to the input which do or do not change the network’s prediction in surprising ways (Szegedy et al., 2014; Jia and Liang, 2017; Ribeiro et al., 2018; Alzantot et al., 2018). Convolutional networks rely heavily on texture (Geirhos et al., 2019), while humans rely on shape (Landau et al., 1988). Neural networks trained to recognize textual entailment can rely heavily on dataset biases (Gururangan et al., 2018).

**Human evaluation setup** We use human evaluation to assess how effectively agents select sentences that also make humans more likely to provide a given answer, when humans act as the judge. Humans answer based only on the question  $Q$ , answer options  $A$ , and a single passage sentence chosen by the agent as evidence for its answer option  $A(i)$  (i.e., using the “Individual Sequential Decision-Making” scheme from §2). Appendix C shows the interface and instructions used to collect evaluations. For each of RACE and DREAM, we use 100 test questions and collect 5 human answers for each  $(Q, A(i))$  pair for each agent. We also evaluate a human baseline for this task, where 3 annotators select the strongest supporting passage sentence for each  $(Q, A(i))$  pair. We report the average results across 3 annotators.

**Humans favor answers supported by evidence agents** when shown that agent’s selected evidence, as shown in Table 3.<sup>1</sup> Without receiving any passage sentences, humans are at ran-

dom chance at selecting the agent’s answer (25% on RACE, 33% on DREAM), since agents are assigned an arbitrary answer. For all evidence agents, humans favor agent-supported answers more often than the baseline (33.5-42.0% on RACE and 41.7-50.5% on DREAM). For our best agents, the relative margin over the baseline is substantial. In fact, these agents select evidence that is comparable to human-selected evidence. For example, on RACE, humans select the target answer 41.6% when provided with human-selected evidence, compared to 42% evidence selected by the learned agent that predicts  $p(i)$ .

All agents support right answers more easily than wrong answers. On RACE, the learned agent that predicts  $p(i)$  finds strong evidence more than twice as often for correct answers than for incorrect ones (74.6% vs. 31.1%). On RACE and DREAM both, BERT-based agents (search or learned agents) find stronger evidence than word-based agents do. Humans tend to find that BERT-based agents select valid evidence for an answer, right or wrong. On DREAM, word-based agents generally fail to find evidence for wrong answers compared to the no-sentence baseline (28.4% vs. 24.5% for a search-based fastText agent).

On RACE, learned agents that predict the  $\text{BERT}_{\text{BASE}}$  judge outperform search agents that directly query the  $\text{BERT}_{\text{BASE}}$  judge. This effect may occur if search agents find an adversarial sentence that unduly affects the judge’s answer but that humans do not find to be valid evidence. Appendix A shows one such example. Learned agents may

<sup>1</sup>Appendix D shows results by question type.

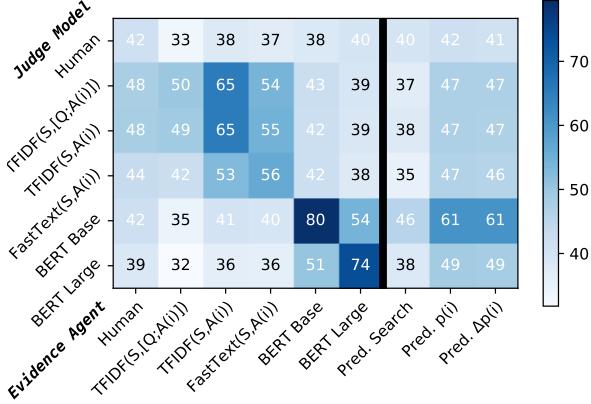


Figure 2: On RACE, how often each judge selects an agent’s answer when given a single agent-chosen sentence. The black line divides learned agents (right) and search agents (left), with human evidence selection in the leftmost column. All agents find evidence that convinces judge models more often than a no-evidence baseline (25%). Learned agents predicting  $p(i)$  or  $\Delta p(i)$  find the most broadly convincing evidence.

have difficulty predicting such sentences, without directly querying the judge. Appendix E provides some analysis on why learned agents may find more general evidence than search agents do. Learned agents are most accurate at predicting evidence sentences when the sentences have a large impact on the judge model’s confidence in the target answer, and such sentences in turn are more likely to be found as strong evidence by humans. On DREAM, search agents and learned agents perform similarly, likely because DREAM has 14x less training data than RACE.

## 4.2 Model Evaluation of Evidence

**Evaluating an agent’s evidence across models**  
 Beyond human evaluation, we test how general agent-selected evidence is, by testing this evidence against various judge models. We expect evidence agents to most frequently convince the model they are optimized to convince, by nature of their direct training or search objective. The more similar models are, the more we expect evidence from one model to be evidence to another. To some extent, we expect different models to rely on similar patterns to answer questions. Thus, evidence agents should sometimes select evidence that transfers to any model. However, we would not expect agent evidence to transfer to other models if models only exploit method-specific patterns.

**Experimental setup** Each agent selects one evidence sentence for each  $(Q, A(i))$  pair. We test how often the judge selects an agent’s answer, when given this sentence,  $Q$ , and  $A$ . We evaluate

on all  $(Q, A(i))$  pairs in RACE’s test set. Human evaluations are on a 100 question subset of test.

**Results** Figure 2 plots how often each judge selects an agent’s answer. Without any evidence, judge models are at random at choosing an agent’s assigned answer (25%). All agents find evidence that convinces judge models more often than the no-evidence baseline. Learned agents that predict  $p(i)$  or  $\Delta p(i)$  find the evidence most broadly considered convincing; other judge models select these agents’ supported answers over 46% of the time. These findings support that evidence agents find general structure despite aiming to convince specific methods with their distinct properties.

Notably, evidence agents are not uniformly convincing across judge models. All evidence agents are most convincing to the judge model they aim to convince; across any given agent’s row, an agent’s target judge model is the model which most frequently selects the agent’s answer. Search agents are particularly effective at finding convincing evidence w.r.t. their target judge model, given that they directly query this model. More broadly, similar models find similar evidence convincing. We find similar results for DREAM (Appendix F).

## 5 Evidence Agents Aid Generalization

We have shown that agents capture method-agnostic evidence representative of answering a question (the strongest evidence for various answers). We hypothesize that QA models can generalize better out of distribution to more challenging questions by exploiting evidence agents’ capability to understand the problem.

Throughout this section, using various train/test splits of RACE, we train a BERT<sub>BASE</sub> judge on easier examples (involving shorter passages or middle-school exams) and test its generalization to harder examples (involving longer passages or high-school exams). Judge training follows §2.1. We compare QA accuracy when the judge answers using (i) the full passage and (ii) only evidence sentences chosen by competing evidence agents. We report results using the round robin competing agent setup described in §2, as it resulted in higher generalization accuracy than free-for-all competition in preliminary experiments. Each competing agent selects sentences up to a fixed, maximum turn limit; we experiment with 3-6 turns per agent (6-12 total sentences for the judge), and we report the best result. We train learned agents (as de-

Train Data	Sentence Selection	<i>RACE → DREAM</i>		
		<i>Sentences in Passage</i>		$\leq 12 \geq 27$
All	Full Passage	64.7	60.0	71.2
<b>RACE</b>	None (Answer-only)	36.1	40.2	38.5
$ S  \leq 12$	Full Passage of Subset	57.4	44.1	65.0
	Random Sentences	49.2	44.7	48.2
	TFIDF( $S, [Q; A(i)]$ )	57.2	48.0	67.3
	fastText( $S, A(i)$ )	<b>57.7</b>	<b>50.2</b>	64.2
	TFIDF( $S, A(i)$ )	57.1	47.9	64.6
	Search over BERT <sub>BASE</sub>	56.7	49.6	<b>68.9</b>
	Predict BERT <sub>BASE</sub> $p(i)$	56.7	50.0	66.9

Table 4: We train a judge on short RACE passages and test its generalization to long passages. The judge is more accurate on long passages when it answers based on only sentences chosen by competing agents (last 5 rows) instead of the full passage. BERT-based agents aid generalization even under test-time domain shift (from RACE to DREAM).

scribed in §2.2) on the full RACE dataset without labels, so these agents can model the judge using more data and on out-of-distribution data.

For reference, we evaluate judge accuracy on a subsequence of randomly sampled sentences; we vary the number of sentences sampled from 6-12 and report the best result. As a lower bound, we train an answer-only model to evaluate how effectively the QA model is using the passage sentences it is given. As an upper bound, we evaluate our BERT<sub>BASE</sub> judge trained on all of RACE, requiring no out-of-distribution generalization.

## 5.1 Generalizing to Longer Passages

We train a judge on RACE passages averaging 10 sentences long (all training passages each with  $\leq 12$  sentences); this data is roughly  $\frac{1}{10}$ th of RACE. We test the judge on RACE passages averaging 30 sentences long.

**Results** Table 4 shows the results. Using the full passage, the judge outperforms an answer-only BERT baseline by 4% (44.1% vs. 40.2%). When answering using the smaller set of agent-chosen sentences, the judge outperforms the baseline by 10% (50.2% vs. 40.2%), more than doubling its relative use of the passage. Both search and learned agents aid the judge model in generalizing to longer passages. The improved generalization is not simply a result of the judge using a shorter passage, as shown by the random sentence selection baseline (44.7%).

Train Data	Sentence Selection	<i>School Level</i>	
		Middle	High
All	Full Passage	70.8	63.2
Middle	None (Answer-only)	38.9	40.2
School only	Full Passage of Subset	66.2	50.7
	Random Sentences	54.8	47.0
	TFIDF( $S, [Q; A(i)]$ )	65.1	50.4
	fastText( $S, A(i)$ )	64.6	50.8
	TFIDF( $S, A(i)$ )	64.9	51.0
	Search over BERT <sub>BASE</sub>	67.0	<b>53.0</b>
	Predict BERT <sub>BASE</sub> $p(i)$	<b>67.3</b>	51.9

Table 5: *Generalizing to harder questions*: We train a judge to answer questions with RACE’s Middle School exam questions only. We test its generalization to High School exam questions. The judge is more accurate when using evidence agent sentences (last 5 rows) rather than the full passage.

## 5.2 Generalizing Across Domains

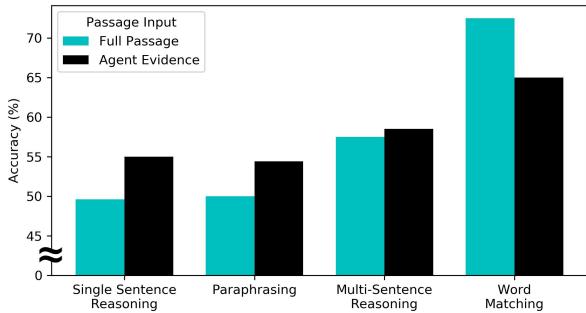
We examine if evidence agents aid generalization even in the face of domain shift. We test the judge trained on short RACE passages on long passages from *DREAM*. We use the same evidence agents from the previous subsection; the learned agent is trained on RACE only, and we do not fine-tune it on DREAM to test its generalization to finding evidence in a new domain. DREAM passages consist entirely of dialogues, use more informal language and shorter sentences, and emphasize general world knowledge and commonsense reasoning (Sun et al., 2019). RACE passages are more formal, written articles (e.g. news or fiction).

**Results** Table 4 shows that BERT-based evidence agents aid generalization even under domain shift. The model shows notable improvements for RACE → DREAM transfer when it predicts from BERT-based agent evidence rather than the full passage (65.0% vs. 68.9%). These results support that our best evidence agents capture something fundamental to the problem of QA, despite changes in e.g. content and writing style.

## 5.3 Generalizing to Harder Questions

Using RACE, we train a judge on middle-school questions and test it on high-school questions.

**Results** Table 5 shows that the judge generalizes to harder questions better by using evidence from either search-based BERT agents (53.0%) or learned BERT agents (51.9%) compared to using the full passage directly (50.7%) or to search-based TFIDF and fastText agents (50.4%-51.0%). Figure 3 shows that the improved generalization comes from questions the model originally gener-



**Figure 3: Generalizing to harder questions by question type:** We train a judge on RACE Middle School questions and test its generalization to RACE High School questions. To predict the answer, the judge uses either the full passage or evidence sentences chosen by a BERT-based search agent. The worse the judge does on a question category using the full passage, the better it does when using the agent-chosen sentences.

alizes worse on. Simplifying the passage by providing key sentences may aid generalization by e.g. removing extraneous or distracting sentences from passages with more uncommon words or complex sentence structure. Such improvements come at the cost of accuracy on easier, word-matching questions, where it may be simpler to answer with the full passage as seen in training.

## 6 Evidence Agents Aid Human QA

As observed in §4.1, evidence agents more easily support right answers than wrong ones. Furthermore, evidence agents do aid QA models in generalizing systematically when all answer evidence sentences are presented at once. We hypothesize that when we combine all evidence sentences, humans prefer to choose the correct answer.

**Human evaluation setup** Evidence agents compete in a free-for-all setup (§2), and the human acts as the judge. We evaluate how accurately humans can answer questions based only on agent sentences. Appendix C shows the annotation interface and instructions. We collect 5 human answers for each of the 100 test questions.

**Humans can answer using evidence sentences alone** Shown in Table 6, humans correctly answer questions using many fewer sentences (3.3 vs. 18.2 on RACE, 2.4 vs. 12.2 on DREAM); they do so while maintaining 90% of human QA accuracy on the full passage (73.2% vs. 82.3% on RACE, 83.8% vs. 93.0% on DREAM). Evidence agents, however, vary in how effectively they aid human QA, compared to answer-agnostic evidence selection. On DREAM, humans answer with 79.1% accuracy using the sentences

Sentences Shown		<b>Human Acc. (%)</b>	
Selection Type	Selection Method	RACE	DREAM
<b>Full Passage</b>	Full Passage	82.3	93.0
<b>No Passage</b>	Answer-only	52.5	43.3
<b>Subset (~20%)</b>	Human Selection	<b>73.5</b>	82.3
<i>Answer-Free Selection</i>	First $n$ Sentences	61.8	68.5
	TFIDF( $S, Q$ )	69.2	77.5
	fastText( $S, Q$ )	69.7	79.1
<i>Search Agent Selection</i>	TFIDF( $S, [Q; A(i)]$ )	66.1	70.0
	TFIDF( $S, A(i)$ )	73.2	77.0
	fastText( $S, A(i)$ )	73.2	77.3
	BERT <sub>BASE</sub>	69.9	<b>83.8</b>
	BERT <sub>LARGE</sub>	72.4	75.0
<i>Learned Agent Selection</i>	Predicting Search	66.5	80.0
	Predicting $p(i)$	71.6	77.8
	Predicting $\Delta p(i)$	65.7	81.5

**Table 6: Human accuracy using evidence agent sentences:** Each agent selects a sentence supporting its own answer. Humans answer the question given these agent-selected passage sentences only. Humans still answer most questions correctly, while reading many fewer passage sentences.

most similar to the question alone (via fastText), while achieving lower accuracy when using the BERT<sub>LARGE</sub> search agent’s evidence (75.0%) and higher accuracy when using the BERT<sub>BASE</sub> search agent’s evidence (83.8%). We explain the discrepancy by examining how effective agents are at supporting right vs. wrong answers (Table 3 from §4.1); BERT<sub>BASE</sub> is more effective than BERT<sub>LARGE</sub> at finding evidence for right answers (82.5% vs. 79.4%) and less effective at finding evidence for wrong answers (34.6% vs. 38.7%).

## 7 Related Work

Here, we discuss further related work, beyond that discussed in §4.1 on (dis)similarities between patterns learned by humans and neural networks.

**Evidence Extraction** Various papers have explored the related problem of extracting evidence or summaries to aid downstream QA. Wang et al. (2018a) concurrently introduced a neural model that extracts evidence specifically for the correct answer, as an intermediate step in a QA pipeline. Prior work uses similar methods to explain what a specific model has learned (Lei et al., 2016; Li et al., 2016; Yu et al., 2019). Others extract evidence to improve downstream QA efficiency over large amounts of text (Choi et al., 2017; Kratzwald and Feuerriegel, 2019; Wang et al., 2018b). More broadly, extracting evidence can facilitate fact verification (Thorne et al., 2018) and debate.<sup>2</sup>

<sup>2</sup>IBM Project Debater: [www.research.ibm.com/artificial-intelligence/project-debater](http://www.research.ibm.com/artificial-intelligence/project-debater)

**Generic Summarization** In contrast, various papers focus primarily on summarization rather than QA, using downstream QA accuracy only as a reward to optimize generic (question-agnostic) summarization models (Arumae and Liu, 2018, 2019; Eyal et al., 2019).

**Debate** Evidence extraction can be viewed as a form of debate, in which multiple agents support different stances (Irving et al., 2018; Irving and Askell, 2019). Chen et al. (2018) show that evidence-based debate improves the accuracy of crowdsourced labels, similar to our work which shows its utility in natural language QA.

## 8 Conclusion

We examined if it was possible to automatically distill general insights for passage-based question answering, by training evidence agents to convince a judge model of any given answer. Humans correctly answer questions while reading only 20% of the sentences in the full passage, showing the potential of our approach for assisting humans in question answering tasks. We examine how selected evidence affects the answers of humans as well as other QA models, and we find that agent-selected evidence is generalizable. We exploit these capabilities by employing evidence agents to facilitate QA models in generalizing to longer passages and out-of-distribution test sets of qualitatively harder questions.

## Acknowledgments

EP was supported by the NSF Graduate Research Fellowship and ONR grant N00014-16-1-2698. KC thanks support from eBay and NVIDIA. We thank Adam Gleave, David Krueger, Geoffrey Irving, Katharina Kann, Nikita Nangia, and Sam Bowman for helpful conversations and feedback. We thank Jack Urbanek, Jason Lee, Ilia Kulikov, Ivanka Perez, Ivy Perez, and our Mechanical Turk workers for help with human evaluations.

## References

- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *EMNLP*, pages 2890–2896.
- Kristjan Arumae and Fei Liu. 2018. Reinforced extractive summarization with question-focused rewards. In *ACL, Student Research Workshop*, pages 105–111.
- Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. In *NAACL*, pages 2566–2577.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.
- Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. 2014. Deep neural networks rival the representation of primate visual cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):1–18.
- David J. Chalmers. 2015. Why isn’t there more progress in philosophy? *Philosophy*, 90(1):331.
- Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Daniel S Weld. 2018. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. *CoRR*, abs/1810.10733.
- Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *ACL*, pages 209–220.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.
- Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. 2018. Feature-wise transformations. *Distill*.
- Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *NAACL*, pages 3938–3948.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*, pages 107–112.
- Geoffrey Irving and Amanda Askell. 2019. AI safety needs social scientists. *Distill*.

- Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. *AI safety via debate*. *CoRR*, abs/1805.00899.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, pages 2021–2031.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *EACL*, pages 427–431.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Bernhard Kratzwald and Stefan Feuerriegel. 2019. Learning from on-line user feedback in neural question answering on the web. In *WWW*, pages 906–916.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReADING Comprehension dataset from Examinations. In *EMNLP*, pages 785–794.
- Barbara Landau, Linda B. Smith, and Susan S. Jones. 1988. The importance of shape in early lexical learning. *Cognitive Development*, 3(3):299–321.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *EMNLP*, pages 107–117.
- Jiwei Li, Will Monroe, and Daniel Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.
- Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *CoRR*, abs/1705.06476.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951.
- Christian S. Perone, Roberto Silveira, and Thomas S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *CoRR*, abs/1806.06259.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, pages 856–865.
- M. Schuster and K. Nakajima. 2012. Japanese and korean voice search. In *ICASSP*, pages 5149–5152.
- Chenglei Si. 2019. Bert for multiple choice machine comprehension.
- Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *EMNLP*, pages 4208–4219.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge dataset and models for dialogue-based reading comprehension. *TACL*, 7:217–231.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *CoRR*, abs/1312.6199.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL*, pages 809–819.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, pages 5998–6008. Curran Associates, Inc.
- Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, Dan Roth, and David McAllester. 2018a. Evidence extraction for machine reading comprehension with deep probabilistic logic. *CoRR*, abs/1902.08852.
- Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2018b. Evidence aggregation for answer re-ranking in open-domain question answering. In *ICLR*.
- Mo Yu, Shiyu Chang, and Tommi S Jaakkola. 2019. Learning corresponded rationales for text matching.