Passage (DREAM)

W: What changes do you think will take place in the next 50 years?
M: I imagine that the greatest change will be the difference between humans and machines.
W: What do you mean?
M: I mean it will be harder to tell the difference between the human and the machine.
W: Can you describe it more clearly?
M: As science develops, it will be possible for all parts of one's body to be replaced. **A computer will work like the human brain.** The computer can recognize one's feelings, and act in a feeling way.
W: You mean man-made human beings will be produced? Come on! That's out of the question!
M: Don't get excited, please. **That's only my personal imagination!**
W: Go on, please. I won't take it seriously.
M: We will then be able to create a machine that is a copy of ourselves. We'll appear to be alive long after we are dead.
W: What a ridiculous idea!
M: It's possible that a way will be found to put our spirit into a new body. Then, we can choose to live as long as we want.
W: In that case, the world would be a hopeless mess!

***Q*: What are the two speakers talking about?**
A. Computers in the future.
B. People's imagination.
C. Possible changes in the future.✓

Table 7: An example from our best evidence agent on DREAM, a search agent using $\text{BERT}_{\text{LARGE}}$. Each evidence agent has chosen a sentence (in color) that convinces a $\text{BERT}_{\text{LARGE}}$ judge model to predict the agent's designated answer with over 99% confidence.

A Additional Evidence Agent Examples

We show additional examples of evidence agent sentence selections in Table 7 (DREAM), as well as Tables 8, 9, and 10 (RACE).

B Implementation Details

B.1 Preprocessing

We use the BERT tokenizer to tokenize the text for all methods (including TFIDF and fastText). To divide the passage into sentences, we use the following tokens as end-of-sentence markers: ".", "?", "!", and the last passage token. For BERT, we use the required WordPiece subword tokenization (Schuster and Nakajima, 2012). For TFIDF, we also use WordPiece tokenization to minimize the number of rare or unknown words. For consistency, this tokenization uses the same vocabulary as our BERT models do. FastText is trained to embed whole words directly, so we do not use subword tokenization.

B.2 Training the Judge

Here we provide additional implementation details of the various judge models.

B.2.1 TFIDF

To limit the number of rare or unknown words, we use subword tokenization via the BERT WordPiece tokenizer. Using this tokenizer enables us to split sentences in an identical manner as for BERT so that results are comparable. For a given dataset, we compute inverse document frequencies for subword tokens using the entire corpus.

B.2.2 BERT

Architecture and Hyperparameters We use the uncased $\text{BERT}_{\text{BASE}}$ pre-trained transformer. We sweep over BERT fine-tuning hyperparameters, using the following ranges: learning rate $\in \{5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$ and batch size $\in \{8, 12, 16, 32\}$.

Segment Embeddings BERT uses segment embeddings to indicate two distinct, contiguous sequences of input text. These segments are also separated by a special `[SEP]` token. The first segment is S, and the second segment is $[Q; A(i)]$.

Truncating Long Passages BERT can only process a maximum of 512 tokens at once. Thus, we truncate the ends of longer passages; we always include the full question Q and answer $A(i)$, as these are generally important in answering the question. We include the maximum number of passage tokens such that the entire input (i.e., (S, Q) or $(S, Q, A(i))$) fits within 512 tokens.

Training Procedure We train for up to 10 epochs, stopping early if validation accuracy decreases after an epoch once (RACE) or 3 times (DREAM). For DREAM, we also decay the learning rate by $\frac{2}{3}$ whenever validation accuracy does not decrease after an epoch.

B.3 Training Evidence Agents

We use the $\text{BERT}_{\text{BASE}}$ architecture for all learned evidence agents. The training details are the same as for the BERT judge, with the exceptions listed below. Agents make sentence-level predictions via end-of-sentence token positions.

Hyperparameters Training learned agents on RACE is expensive, due to the dataset size and number of answer options to make predictions for. Thus, for these agents only (not DREAM agents),

Passage (RACE)
Who doesn't love sitting beside a cosy fire on a cold winter's night? Who doesn't love to watch flames curling up a chimney? Fire is one of man's greatest friends, but also one of his greatest enemies. **Many big fires are caused by carelessness.** **A lighted cigarette thrown out of a car or train window or a broken bottle lying on dry grass can start a fire.** **Sometimes, though, a fire can start on its own.** Wet hay can begin burning by itself. This is how it happens: the hay starts to rot and begins to give off heat which is trapped inside it. Finally, it bursts into flames. **That's why farmers cut and store their hay when it's dry.** Fires have destroyed whole cities. In the 17th century, a small fire which began in a baker's shop burnt down nearly every building in London. Moscow was set on fire during the war against Napoleon. This fire continued burning for seven days. And, of course, in 64 A.D. a fire burnt Rome. Even today, in spite of modern fire-fighting methods, fire causes millions of pounds' worth of damage each year both in our cities and in the countryside. It has been wisely said that fire is a good servant but a bad master.
Q: Many big fires are caused... **A.** by cigarette **B.** by their own **C.** by dry grass **D.** by people's carelessness ✓

Table 8: In this example, each answer's agent has chosen a sentence (in color) that individually influenced a neural QA model to answer in its favor. When human evaluators answer the question using only one agent's sentence, evaluators select the agent-supported answer. When humans read all 4 agent-chosen sentences together, they correctly answer "D", without reading the full passage.

Passage (RACE)
Yueyang Tower lies in the west of Yueyang City, near the Dongting Lake. It was first built for soldiers to rest on and watch out. In the Three Kingdoms Period, Lu Su, General of Wu State, trained his soldiers here. **In 716, Kaiyuan of Tang Dynasty, General Zhang Shuo was sent to defend at Yuezhou and he rebuilt it into a tower named South Tower, and then Yueyang Tower.** **In 1044, Song Dynasty, Teng Zijing was stationed at Baling Jun, the ancient name of Yueyang City.** In the second year, he had the Yueyang Tower repaired and had poems by famous poets written on the walls of the tower. Fan Zhongyan, a great artist and poet, was invited to write the well - known poem about Yueyang Tower. **In his *A Panegyric of the Yueyang Tower*, Fan writes: "Be the first to worry about the troubles across the land, the last to enjoy universal happiness."** His words have been well - known for thousands of years and made the tower even better known than before. The style of Yueyang Tower is quite special. The main tower is 21.35 meters high with 3 stories, flying eave and wood construction, the helmet-roof of such a large size is a rarity among the ancient architectures in China. **Entering the tower, you'll see "Dongting is the water of the world, Yueyang is the tower of the world".** Moving on, there is a platform that once used as the training ground for the navy of Three-Kingdom Period general Lu Su. To its south is the Huaifu Pavilion in honor of Du Fu. Stepping out of the Xiaoxiang Door, the Xianmei Pavilion and the Sanzui Pavilion can be seen standing on two sides. In the garden to the north of the tower is the tomb of Xiaoqiao, the wife of Zhou Yu.
Q: Yueyang Tower was once named... **A.** South Tower ✓ **B.** Xianmei Tower **C.** Sanzui Tower **D.** Baling Tower

Table 9: An example where each answer's search agents successfully influences the answerer to predict that agent's answer; however, the supporting sentence for "B" and for "C" are not evidence for the corresponding answer. These search agents have found adversarial examples in the passage that unduly influence the answerer. Thus, it can help to present the answerer model with evidence for 2+ answers at once, so the model can weigh potentially adversarial evidence against valid evidence. In this case, the model correctly answers "B" when predicting based on all 4 agent-chosen sentences.

Passage (RACE)
A desert is a beautiful land of silence and space. **The sun shines, the wind blows, and time and space seem endless.** Nothing is soft. The sand and rocks are hard, and many of the plants even have hard needles instead of leaves. **The size and location of the world's deserts are always changing.** Over millions of years, as climates change and mountains rise, new dry and wet areas develop. But within the last 100 yeas, deserts have been growing at a frightening speed. This is partly because of natural changes, but the greatest makers are humans. **Humans can make deserts, but humans can also prevent their growth.** **Algeria Mauritania is planting a similar wall around Nouakchott, the capital.** Iran puts a thin covering of petroleum on sandy areas and plants trees. The oil keeps the water and small trees in the land, and men on motorcycles keep the sheep and goats away. The USSR and India are building long canals to bring water to desert areas.
Q: Which of the following is NOT true? **A.** The greatest desert makers are humans. **B.** There aren't any living things in the deserts. ✓ **C.** Deserts have been growing quickly. **D.** The size of the deserts is always changing.

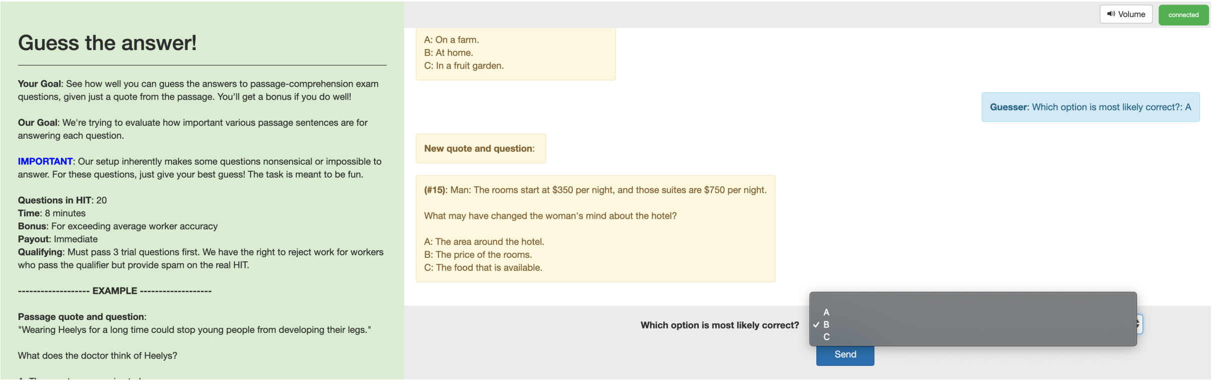
Table 10: In this example, the answerer correctly predicts "B," no matter the passage sentence (in color) a search agent provides. This behavior occurred in several cases where the question and answer options contained a strong bias in wording that cues the right answer. Statements including "all," "never," or "there aren't any" are often false, which in this example signals the right answer. Gururangan et al. (2018) find similar patterns in natural language inference data, where "no," "never," and "nothing" strongly signal that one statement contradicts another.

we sweep over a limited range that works well: learning rate $\in \{5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$ and batch size $\in \{12\}$.

Training Procedure We use early stopping based on validation loss instead of answering accuracy, since evidence agents do not predict the correct answer.

C Human Evaluation Details

For all human evaluations, we filter out workers who perform poorly on a few representative examples of the evaluation task. We pay workers on average $15.48 per hour according to TurkerView (`https://turkerview.com`). We require workers to be from predominantly English-speaking countries: Australia, Canada, Great Britain, New Zealand, or the U.S.



Guess the answer!

Your Goal: See how well you can guess the answers to passage-comprehension exam questions, given just a quote from the passage. You'll get a bonus if you do well!

Our Goal: We're trying to evaluate how important various passage sentences are for answering each question.

IMPORTANT: Our setup inherently makes some questions nonsensical or impossible to answer. For these questions, just give your best guess! The task is meant to be fun.

Questions in HIT: 20
Time: 8 minutes
Bonus: For exceeding average worker accuracy
Payout: Immediate
Qualifying: Must pass 3 trial questions first. We have the right to reject work for workers who pass the qualifier but provide spam on the real HIT.

------------------- EXAMPLE -------------------

Passage quote and question:
"Wearing Heelys for a long time could stop young people from developing their legs."

What does the doctor think of Heelys?

A: On a farm.
B: At home.
C: In a fruit garden.

Guesser: Which option is most likely correct?: A

New quote and question:

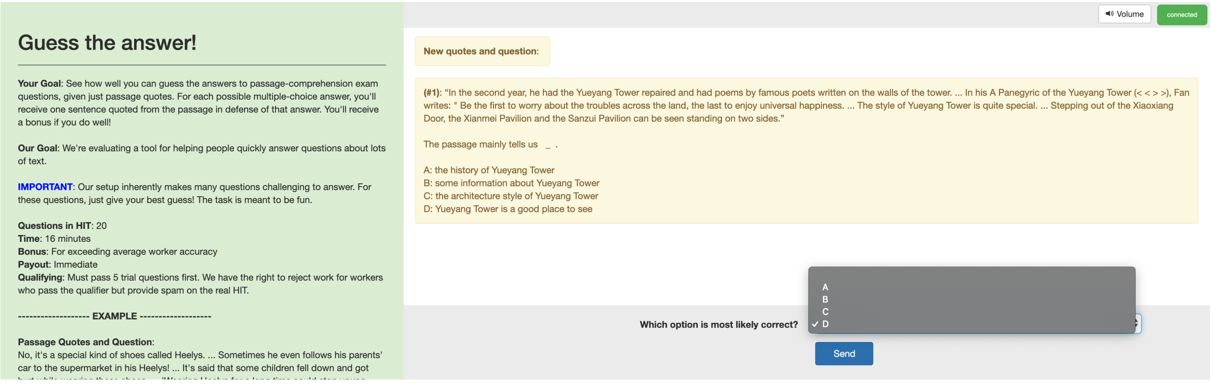
(#15): Man: The rooms start at $350 per night, and those suites are $750 per night.

What may have changed the woman's mind about the hotel?

A: The area around the hotel.
B: The price of the rooms.
C: The food that is available.

Which option is most likely correct?
A
✓ B
C

Figure 4: Interface for humans to answer questions based on one agent-selected passage sentence only. In this example from DREAM, a learned agent supports the correct answer (B).



Guess the answer!

Your Goal: See how well you can guess the answers to passage-comprehension exam questions, given just passage quotes. For each possible multiple-choice answer, you'll receive one sentence quoted from the passage in defense of that answer. You'll receive a bonus if you do well!

Our Goal: We're evaluating a tool for helping people quickly answer questions about lots of text.

IMPORTANT: Our setup inherently makes many questions challenging to answer. For these questions, just give your best guess! The task is meant to be fun.

Questions in HIT: 20
Time: 16 minutes
Bonus: For exceeding average worker accuracy
Payout: Immediate
Qualifying: Must pass 5 trial questions first. We have the right to reject work for workers who pass the qualifier but provide spam on the real HIT.

------------------- EXAMPLE -------------------

Passage Quotes and Question:
No, it's a special kind of shoes called Heelys. ... Sometimes he even follows his parents' car to the supermarket in his Heelys! ... It's said that some children fell down and got

New quotes and question:

(#1): “In the second year, he had the Yueyang Tower repaired and had poems by famous poets written on the walls of the tower. ... In his A Panegyric of the Yueyang Tower (< < > >), Fan writes: " Be the first to worry about the troubles across the land, the last to enjoy universal happiness. ... The style of Yueyang Tower is quite special. ... Stepping out of the Xiaoxiang Door, the Xianmei Pavilion and the Sanzui Pavilion can be seen standing on two sides.”

The passage mainly tells us _ .

A: the history of Yueyang Tower
B: some information about Yueyang Tower
C: the architecture style of Yueyang Tower
D: Yueyang Tower is a good place to see

Which option is most likely correct?
A
B
C
✓ D

Figure 5: Interface for humans to answer questions based on agent-selected passage sentences only. Each answer's evidence agent selects one sentence. These sentences are combined and shown to the human, in the order they appear in the passage. In this example from RACE, the agents are search-based, and the correct answer is B.

We do not use results from workers who complete the evaluation significantly faster than other workers (i.e., less than a few seconds per question). To incentivize workers, we also offer a bonus for answering questions more accurately than the average worker. Figures 4 and 5 show two examples of our evaluation setup.

D Human Evaluation of Agent Evidence by Question Category

We show a detailed breakdown of results from §4.1, where humans answer questions using an agent-chosen sentence. Table 11 shows how often humans select the agent-supported answer, broken down by question type. Models that perform better generally do so across all categories. However, methods incorporating neural methods generally achieve larger gains over word-based methods on multi-sentence reasoning questions on RACE.

E Analysis

Highly convincing evidence is easiest to predict Figure 6 plots the accuracy of a search-predicting evidence agent at predicting the search-chosen sentence, based on the magnitude of that sentence's effect on the judge's probability of the target answer. Search-predicting agents more easily predict search's sentence the greater the effect that sentence has on the judge's confidence.

Strong evidence to a model tends to be strong evidence to humans as shown in Figure 7. Combined with the previous result, we can see that learned agents are more accurate at predicting sentences that humans find to be strong evidence.

F Model Evaluation of Evidence on DREAM

Figure 8 shows how convincing various judge models find each evidence agent. Our findings on DREAM are similar to those from RACE in §4.2.

		How Often Human Selects Agent's Answer (%)												
		RACE								***DREAM***				
			School Level		*Question Type*						*Question Type*			
	Evidence Sentence Selection Method	Overall	Middle	High	Word Match	Para-phrase	Single Sent. Reasoning	Multi-Sent. Reasoning	Ambi-guous	Overall	Common Sense	Logic	Word-Match/ Paraphrase	Summary
Baselines	No Sentence	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	33.3	33.3	33.3	33.3	33.3
	Human Selection	38.1	46.4	39.5	44.6	41.3	41.7	41.7	38.5	50.7	50.0	50.6	48.2	52.1
Search Agents querying...	TFIDF$(S, [Q; A(i)])$	33.5	36.5	32.2	35.0	36.1	31.8	34.2	32.7	41.7	37.2	42.4	37.1	41.8
	TFIDF$(S, A(i))$	38.0	41.8	36.4	44.8	39.9	38.4	35.2	31.1	43.4	40.0	42.7	46.4	42.7
	fastText$(S, A(i))$	37.1	40.3	35.7	38.2	37.9	38.1	36.2	34.4	41.5	41.0	42.2	37.0	40.7
	BERT$_{\text{BASE}}$	38.4	40.4	37.5	44.5	36.7	39.2	37.2	39.4	50.5	48.2	**50.6**	52.1	50.2
	BERT$_{\text{LARGE}}$	40.1	44.5	38.3	41.3	38.8	39.9	**42.0**	39.0	**52.3**	**49.8**	50.3	**59.3**	**54.5**
Learned Agents: **predicting...**	Search	40.0	42.0	39.2	43.7	41.8	39.3	41.2	38.1	49.1	44.6	49.9	47.9	45.9
	$p(i)$	**42.0**	44.3	**41.0**	**47.0**	**43.6**	**42.3**	41.9	34.3	50.0	47.6	50.1	47.3	49.6
	$\Delta p(i)$	41.1	**44.9**	39.5	43.7	41.4	41.0	41.9	**39.6**	48.2	45.5	47.1	55.5	47.2

Table 11: *Human evaluations*: **Search Agents** select evidence by querying the specified judge model, and **Learned Agents** predict the strongest evidence w.r.t. a judge model (BERT$_{\text{BASE}}$); humans then answer the question using the selected evidence sentence (without the full passage).

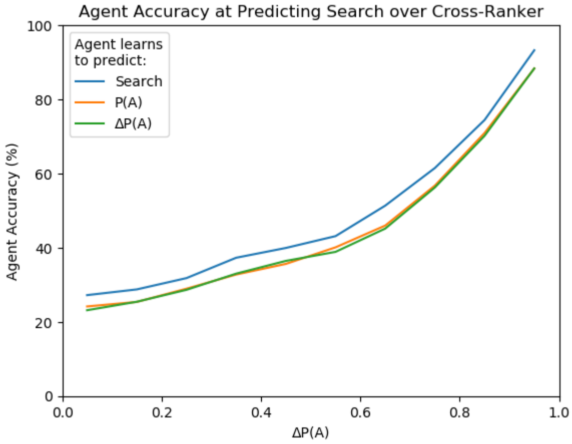


Figure 6: Learned agent validation accuracy at predicting the top sentence chosen by search over the judge (BERT$_{\text{BASE}}$ on RACE). The stronger evidence a judge model finds a sentence to be, the easier it is to predict as the being an answer's strongest evidence sentence in the passage. This effect holds regardless of the agent's particular training objective.

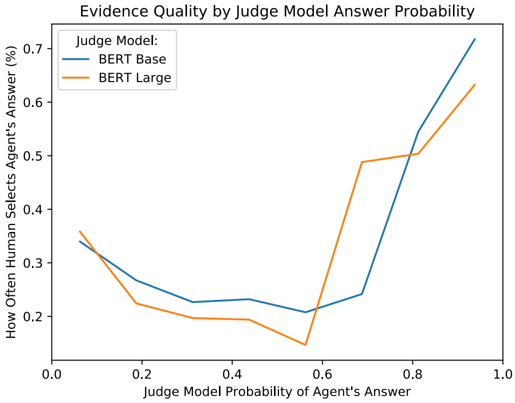


Figure 7: We find the passage sentence that would best support an answer to a particular judge model (i.e., using a search agent). We plot the judge's probability of the target answer given that sentence against how often humans also select that target answer given that same sentence. Humans tend to find a sentence to be strong evidence for an answer when the judge model finds it to be strong evidence.

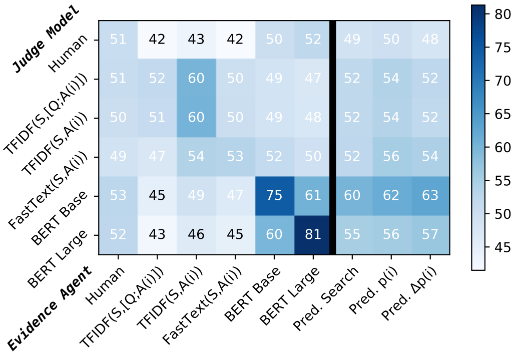


Figure 8: On DREAM, how often each judge selects an agent's answer when given a single agent-chosen sentence. The black line divides learned agents (right) and search agents (left), with human evidence selection in the leftmost column. All agents find evidence that convinces judge models more often than a no-evidence baseline (33%). Learned agents predicting $p(i)$ or $\Delta p(i)$ find the most broadly convincing evidence.